FINAL REPORT

Supervised Classification of Underwater Optical Imagery for Improved Detection and Characterization of Underwater Military Munitions

SERDP Project MR-2414

JUNE 2015

Arthur Gleason
Brooke Gintert
University of Miami

Nuno Gracias
A. Shihavuddin
University of Girona

Greg Schultz
White River Technologies

SERDP

DOD ▪ EPA ▪ DOE

This report was prepared under contract to the Department of Defense Strategic Environmental Research and Development Program (SERDP). The publication of this report does not indicate endorsement by the Department of Defense, nor should the contents be construed as reflecting the official policy or position of the Department of Defense. Reference herein to any specific commercial product, process, or service by trade name, trademark, manufacturer, or otherwise, does not necessarily constitute or imply its endorsement, recommendation, or favoring by the Department of Defense.

REPORT DOCUMENTATION PAGE	*Form Approved* *OMB No. 0704-0188*

1. REPORT DATE *(DD-MM-YYYY)* 31-05-2015	**2. REPORT TYPE** Final	**3. DATES COVERED** *(From - To)* April 2014 - May 2015

4. TITLE AND SUBTITLE Supervised Classification of Underwater Optical Imagery for Improved Detection and Characterization Of Underwater Military Munitions	**5a. CONTRACT NUMBER** W912HQ-14-P-0012
	5b. GRANT NUMBER
	5c. PROGRAM ELEMENT NUMBER
6. AUTHOR(S) Gleason, Arthur C.R. Gintert, Brooke Gracias, Nuno Shihavuddin, Asm Schultz, Gregory	**5d. PROJECT NUMBER** MR-2414
	5e. TASK NUMBER
	5f. WORK UNIT NUMBER
7. PERFORMING ORGANIZATION NAME(S) AND ADDRESS(ES) University of Miami 1551 Brecia Ave. Rm 100A Coral Gables, FL 33146-2503	**8. PERFORMING ORGANIZATION REPORT NUMBER**
9. SPONSORING / MONITORING AGENCY NAME(S) AND ADDRESS(ES) Strategic Environmental Research and Development Program 4800 Mark Center Drive Suite 17D08 Alexandria VA 22350-3600	**10. SPONSOR/MONITOR'S ACRONYM(S)** SERDP
	11. SPONSOR/MONITOR'S REPORT NUMBER(S) MR-2414

12. DISTRIBUTION / AVAILABILITY STATEMENT

Approved for public release; distribution is unlimited

13. SUPPLEMENTARY NOTES

N/A

14. ABSTRACT

Optical images of the seabed could benefit surveys for underwater military munitions (UWMM). Due to the need for human interpretation, however, analysis is currently a bottleneck for quantitative assessment of underwater images. In this project, a recently developed image classification algorithm was tested for identifying UWMM and seabed types. Also, an extension to the algorithm using seabed microtopography, or roughness, features was developed and tested. The image classifier by itself distinguished munitions from non-munitions (background) with generally > 80% accuracy. Discrimination of environments was high for the major seabed types. For example, sand and mixed sand-seagrass were classified with 80-100% accuracy in both shallow and deep water. Extending the algorithm to also use height data derived from stereo reconstruction greatly improved the classification results. Improved accuracy with the height features was observed not only on the basic, binary munitions / non-munitions classes, but also improved the capability to discriminate different types of munitions from one another.

15. SUBJECT TERMS

underwater military munitions, seabed classification, 2.5-D features

16. SECURITY CLASSIFICATION OF:			**17. LIMITATION OF ABSTRACT**	**18. NUMBER OF PAGES**	**19a. NAME OF RESPONSIBLE PERSON** Arthur Gleason
a. REPORT U	**b. ABSTRACT** U	**c. THIS PAGE** U	SAR	41	**19b. TELEPHONE NUMBER** *(include area code)* (305) 284-7140

Table of Contents

List of Tables

List of Figures

List of Acronyms

2-D: two-dimensional
2.5-D: two and a half-dimensional
3-D: three dimensional
AUV: autonomous underwater vehicle
CLBP: completed local binary pattern

DEM: digital elevation model
ESTCP
GLCM: grey level co-occurrence matrix
KNN: k-nearest neighbor
HDV: high-definition video
HUMMA: Hawaii Undersea Military Munitions Assessment
NOAA: National Oceanic and Atmospheric Administration
OA: overall accuracy
PCA: principal components analysis
PDWMD: probability density weighted mean distance
ROI: regions of interest
ROV: remotely operated vehicle
SERDP
SfM: structure from motion
SIP: spatial interest pixel
SON: statement of need
SVM: support vector machine
TFCM: texture feature coding method
USA: United States of America
UWMM: underwater military munitions
UXO: unexploded ordnance
WHOI: Woods Hole Oceanographic Institution

Keywords
underwater military munitions, seabed classification, 2.5-D features

Acknowledgements

We gratefully thank Margo Edwards at the University of Hawaii, who contributed to this project by providing access to a large existing database of seabed images acquired south of Pearl Harbor, Hawaii, with the support of the U.S. Army and other U.S. federal partners, as part of the Hawaii Undersea Military Munitions Assessment (HUMMA) project. Pamela Reid, at the University of Miami, provided equipment, laboratory space, and thoughtful comments on drafts of this report. The efforts of Chris Kelley, Geoff Carton, and Michael Tomlinson, who identified munitions in the HUMAA data, are much appreciated.

1 Abstract:

1.1 Objectives

The Department of Defense needs cost-effective methods for locating and identifying navigation and safety hazards related to underwater military munitions (UWMM). This is a broad and diverse problem because historical information relating to the locations of underwater munitions is often limited and not always accurate. Furthermore, some munitions were dispersed over large areas, being dumped by vessels, dropped via aircraft, or shot as projectiles on or adjacent to live target ranges.

Technologies are needed that can efficiently and objectively detect, identify, and map underwater military munitions. Furthermore, knowledge of benthic environments adjacent to UWMM is critical for remediation decisions. Managers need to know, for example, if detected munitions and explosives of concern are encrusted in a reef or are partially buried; are they intact or cracked open; are they mobile or hard-stuck in silty mud? Currently, both wide-area searches and detailed mapping of underwater military munitions rely on multiple acoustic (*e.g.* side-scan or multi-beam sonar) and/or metal detection methods.

Underwater optical images of the seabed are another technology that could benefit surveys for UWMM as well as efforts to understand the environments around UWMM because optical images have very high spatial resolution. Objects on the seabed as small as a few cm in diameter can be easily resolved with optical imagery. In many cases, the type or condition of munitions, such as whether they are intact or not, can be discerned as well. Currently, however, analysis is a bottleneck for *quantitative* assessment of underwater images; therefore imagery tends to be used for UWMM response in only a *qualitative* way involving visual inspection and interpretation by an analyst.

Our group recently developed a new seabed classification algorithm that has been shown to accurately classify benthic images from coral reefs (Shihavuddin et al. 2013). Three technical objectives were identified to evaluate the potential of underwater seabed images to improve both wide-area and detailed surveys for UWMM. First, how well can the algorithm classify munitions targets? Second, how well can the algorithm classify seabed types, thereby characterizing the environments around munitions targets? Third, would further development on the algorithm yield improvements in the ability to discriminate munitions and aspects of the environment?

1.2 Technical Approach

Prior to this project, the supervised classification algorithm we had developed had been tested with images of coral reefs from three different locations, and one dataset of seabed images from a location where UWMM were found. Based on positive results from these preliminary tests, the approach for this project was to use two large datasets to test the algorithm in a wider variety of settings. The first dataset was collected specifically for this project, using inert surrogate munitions placed on the seabed as known targets. The second set of tests exploited an existing dataset of over 30,000 images collected south of Pearl Harbor, HI, covering an area where actual chemical and conventional munitions were discarded over a 50-year period. Together, these datasets will provide an assessment of the capabilities of the current classification algorithm across a wide spectrum of depths, substrates, benthic biota, and munitions types.

In addition to tests of the existing supervised classification algorithm, we investigated the utility of adding additional features to the classifier that were based on the local relief, or height, of the seabed. Height data was generated from the input images themselves using structure-from-motion computer vision techniques. A third dataset, also from Hawaii, was used to develop the modified algorithm, which was then tested on the newly collected dataset from Miami.

1.3 Results

The image classifier by itself was shown to distinguish munitions from non-munitions (background) with generally high (> 80%) accuracy. This was accomplished at multiple sites in shallow water over seagrass, reef, and sand, and at depths greater than 500 m in sand. Discrimination of environments was high for the major seabed types. For example, sand and mixed sand-seagrass were classified with 80-100% accuracy in both shallow and deep water.

The image classifier by itself did not achieve > 80% accuracy for every class in every situation, however, even with a binary munitions/non-munitions scheme. False positive matches for munitions were observed in the Miami dataset over reef and seagrass, and false negatives were observed in the Hawaii dataset due to confusion of munitions with other anthropogenic clutter. These limitations indicated that, indeed, there remains a need for improvements to the algorithm.

Extending our existing algorithm to also use height data derived from stereo reconstruction showed that incorporating such so-called “2.5-D” data greatly improved the classification results. Using the 2.5-D information reduced the number of false positives in the Miami dataset. Furthermore, improved accuracy was observed not only on the basic, binary munitions / non-munitions classes. Adding 2.5-D information improved the capability to discriminate different types of munitions from one another.

1.4 Benefits

UWMM response would be enhanced by adding quantitative image analysis to the toolbox of survey methodologies. Furthermore, any other application requiring quantitative assessment of benthic communities would also benefit from the same tools. Thus, an accurate, automated algorithm that can classify seabed images from diverse environments would be a benefit both to the Department of Defense and the wider scientific community.

2 Objective

2.1 SERDP Relevance

This project tested the use of supervised classification of underwater optical imagery for the detection and characterization of underwater military munitions (UWMM). Optical images of the seabed have the potential to improve *both* wide-area and detailed surveys for UWMM. Furthermore, underwater images also have the potential to characterize UWMM as well as the surrounding benthic biota and substrate. These were all areas of interest according to the statement of need (SON) MRSEED-14-01.

The strength of underwater images of the seabed is their high spatial resolution, which is one to two orders of magnitude greater than other survey technologies. The weakness of these images, historically, has been the labor involved to extract quantitative information from them. Thus, optical imagery is currently used only as an inspection tool and not as a survey methodology for UWMM response (Schwartz and Brandenburg 2009). Other communities, such as benthic ecologists, do use seabed images quantitatively, but the quantitative information is extracted using laborious, manual methods such as point counting (Kohler and Gill 2006).

An automated process for classifying underwater optical imagery would reduce the analysis bottleneck and allow large underwater image datasets to be fully exploited for UWMM surveys. Efforts to automate underwater image classification have been made for well over a decade (Pican et al. 1998), but no single technique is yet widely accepted as robust. Our group recently developed a new seabed classification algorithm that has been shown on a handful of datasets to accurately classify benthic images from coral reefs (Shihavuddin et al. 2013).

Evaluation of our new seabed-image classification algorithm is relevant to the SERDP SEED program for three reasons. First, the use of images for quantitative UWMM surveys is uncommon and therefore unproven or “risky.” Second, diverse pilot datasets are needed to test the approach. Third, if initial tests on new datasets are positive, we see a clear path forward for future development that is expected to improve the algorithm’s capabilities.

2.2 Technical Objectives

The overall objective of this proposal was to assess the potential of using underwater images of the seabed for the detection of UWMM and for characterizing the surrounding environment. Three technical objectives were the logical first step toward a long-term goal of fully integrating seabed imagery with other sensor types for UWMM surveys.

Objective 1: Evaluate the potential for automated classification of underwater seabed images to improve both wide-area and detailed surveys for UWMM.

Objective 2: Evaluate the potential for automated classification of underwater seabed images to improve characterization of the environment surrounding UWMM by discriminating bottom types in general (rock, sediment, seagrass etc..) as well as identifying specific coral species.

Objective 3: Test the assumption that further improvements to the current generation of seabed classification algorithms will yield further improvements in the ability to discriminate munitions and aspects of the environment, such as coral species identification.

The technical objectives of this proposal were a logical proof of concept step toward the long-term goal of this work, which is to incorporate optics into a multi-modal approach to UWMM surveys. In such an approach, seabed images would be used along with acoustics, magnetometers, and active source electromagnetic arrays. The rationale for this approach is that each sensing modality has its strengths and weaknesses, so using multiple techniques together should result in improvements over any single technology (SERDP and ESTCP 2007).

High spatial resolution is the primary strength of optics. Images of the seabed can be acquired with cm to sub-mm spatial resolution, which is sufficient to accomplish many of the goals of the statement of need (SON). It's clear that an analyst can *visually* distinguish different types of munitions, assess the condition of munitions, discriminate munitions from clutter, assess the surficial environment around munitions, and identify benthic biota. The research question that needs to be addressed is whether these processes can be *automated* so that wide areas of investigation can be processed efficiently and incorporated into geospatial data products.

The technical objectives that are needed to achieve the long-term goal, therefore, are a) to develop classification methods that fuse optical images with other co-located data sources such as acoustics and magnetics, and b) to improve automated seabed classification from optical seabed images (Fig. 1). The starting point for this goal was our existing seabed classification algorithm (Shihavuddin et al. 2013). In the long term (gray box in Fig. 1), the approach for data fusion will start with the existing optical algorithm and combine it with classification methods for complementary acoustic and magnetic datasets. Likewise, in the long term, the approach for automated seabed classification will be to improve on the existing optical algorithm by incorporating new features based on shape and scale.

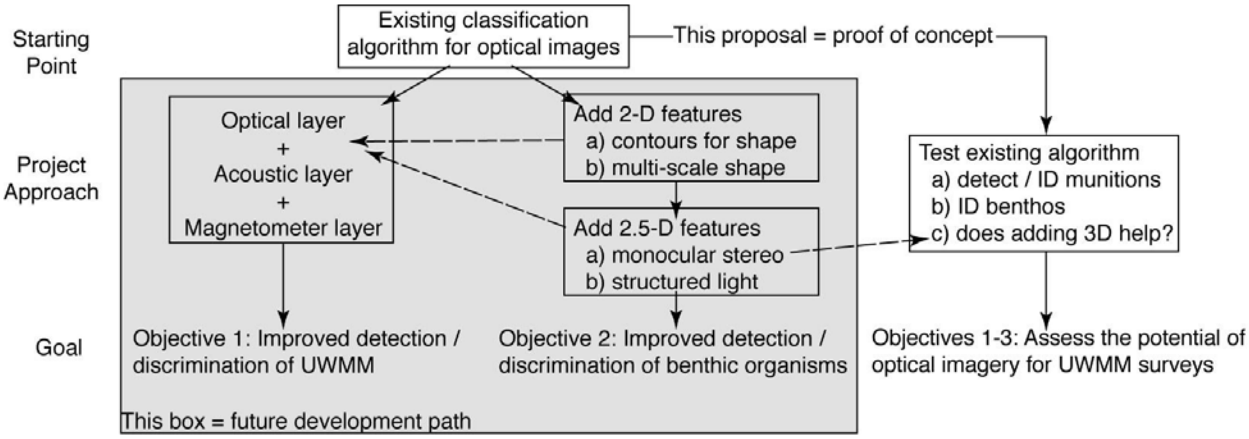

Starting Point
Existing classification algorithm for optical images
This proposal = proof of concept
Project Approach
Optical layer + Acoustic layer + Magnetometer layer
Add 2-D features
a) contours for shape
b) multi-scale shape
Add 2.5-D features
a) monocular stereo
b) structured light
Test existing algorithm
a) detect / ID munitions
b) ID benthos
c) does adding 3D help?
Goal
Objective 1: Improved detection / discrimination of UWMM
Objective 2: Improved detection / discrimination of benthic organisms
Objectives 1-3: Assess the potential of optical imagery for UWMM surveys
This box = future development path

Figure 1. Project objectives in the context of long-term research goals: The long-term goal of using optical imagery of the seabed to improve the detection of UWMM and improve characterization of the environment surrounding UWMM requires additional development. The gray box indicates the future development path for a multi-modal approach to UWMM surveys. As a proof-of-concept, this project tested how well our existing algorithm can work on its own, and explored the degree to which monocular stereo features, which is one of the envisioned future enhancements, could improve on the existing algorithm.

In the short term, this proof of concept study will reduce the risk of investing in extensive algorithm improvement by demonstrating the potential of the existing algorithm, and the potential benefits of improving the existing algorithm (Fig. 1 not including gray box).

Background

Underwater military munitions have generally received less attention than munitions on land because of the perceived lower risks of human exposure as well as the cost of detecting, locating, classifying, and removing underwater munitions compared to remediation on land. Furthermore, historical information relating to the locations of underwater munitions is often limited and not always accurate. Some munitions were dispersed over large areas, being dumped by vessels, dropped via aircraft, or shot as projectiles on or adjacent to live target ranges. Nevertheless, in response to requirements in the John Warner National Defense Authorization Act for FY 2007, 109-364, Section 314 (b)(1), the Department of Defense needs cost-effective methods for locating and identifying navigation and safety hazards related to UWMM and must provide that information to the Secretary of Commerce to assist the National Oceanic and Atmospheric Administration in preparing nautical charts and other navigational materials for coastal waters.

On land, a multi-modal approach (*i.e.* one in which multiple survey technologies are used) to discarded munitions mapping has proven successful (Nelson et al. 2008; Foley and Hodgson 2010; Keranen and O'Neil-Dunne 2010). Underwater, different tools must be used, but a multi-modal approach is expected to have similar success (Schultz et al. 2009). No single technology works best in all situations. A layered approach using several different survey technologies provides the most robust solution.

High spatial resolution is the primary advantage to including images of the seabed in UWMM surveys. Objects on the seabed as small as a few cm in diameter can be easily resolved with optical imagery (Fig. 2). In many cases, the type or condition of munitions, such as whether they are intact or not, can be discerned as well (Fig. 2B, 2C). Currently, however, both wide-area searches and detailed mapping of UWMM rely on multiple acoustic (*e.g.* side-scan or multi-beam sonar) and/or metal detection methods (SERDP and ESTCP 2007; Schwartz and Brandenburg 2009). Images tend to only be used for UWMM response in a qualitative way, for visual inspection and interpretation by an analyst (Schwartz and Brandenburg 2009).

One reason imagery has not been widely used for quantitative UWMM mapping is that optics are not appropriate in every situation. Nevertheless, imagery is a critical asset in shallow water with good visibility and exposed munitions on the seafloor. Such locations are high-priority areas for UWMM management. Locations with good visibility and shallow depths (*i.e.* < 40 m recreational SCUBA limit) are favored sites for human recreation. Rocky seabeds, such as coral or bedrock reefs, where munitions are likely to be exposed, are also areas that attract fishing and diving activity. When shallow, rocky areas with good visibility contain munitions, public use for recreational boating, diving, snorkeling, and swimming increases risk of human exposure and public concern. Furthermore, rocky seabeds are also the areas where traditional magnetometer and electromagnetic geophysical surveys are least effective due to increased standoff requirements in order to avoid colliding with the bottom. Additionally, new methods of obtaining imagery in limited visibility are currently being developed (Schechner and Karpel 2005) that should expand the utility of this method.

A second reason imagery has not been widely used for quantitative UWMM mapping is that, until recently, obtaining data over large areas was difficult due to the small footprint of individual images. This obstacle has largely been eliminated in the past decade with the widespread availability of digital cameras, associated advancements in batteries and storage, and increased sophistication and availability of platforms such as remotely operated vehicles (ROVs), autonomous underwater vehicles (AUVs), and towed bodies. Additionally, images can be stitched together to form mosaics of the seabed covering 100's to 1000's of m^2 {Lirman, 2007 #593 and Fig. 2}. It is now easy and common to collect more imagery than can be practically analyzed by hand.

The third reason imagery has not been widely used for quantitative UWMM mapping is that it has traditionally required laborious manual analysis. Software packages such as CPCe (Kohler and Gill 2006) have facilitated the extraction of quantitative information from imagery, but humans are still "in the loop." The full potential of optical imagery for seabed mapping will not be realized until automated classification is implemented, hence the need for research into these methods.

Efforts to automate classification of the seabed using optical imagery have been made for well over a decade (Pican et al. 1998; Marcos et al. 2008; Pizarro et al. 2008; Stokes and Deane 2009; Beijbom et al. 2012) but no single algorithm is yet widely accepted as robust. A few of the most challenging obstacles to classification accuracy in natural environments include: significant intra-class and inter-site variability in the morphology of benthic organisms, complex spatial borders between classes on the seabed, subjective annotation of training data by different analysts, variation in viewpoints, distances, and image quality, limits to spatial and spectral resolution when trying to classify to a free taxonomic scale, partial occlusion of objects due to the three-dimensional structure of the seabed, lighting artifacts due to wave focusing, and variable optical properties of the water column.

To overcome some of these challenges, a seabed classification algorithm for optical images was recently developed by our team. The algorithm uses a feature vector computed from completed local binary pattern (CLBP), grey level co-occurrence matrix (GLCM), Gabor filter response, and color histograms. Either k-nearest neighbor (KNN), support vector machine (SVM) or probability density weighted mean distance (PDWMD) classifiers can be used, depending on the characteristics of the dataset (Shihavuddin et al. 2013). The algorithm has been tested on datasets from Florida, USA, Moorea, French Polynesia, and the Red Sea with overall accuracy of 96% on 14 classes, 85% on 9 classes, and 97% on 8 classes, respectively (Shihavuddin et al. 2013). The method uses both texture and color, and, based on the datasets previously tested, was a good starting point for discrimination of UWMM.

Our objectives 1 and 2 followed directly from the results obtained thus far from our seabed classification algorithm (Shihavuddin et al. 2013). The next step for evaluating its potential use in UWMM surveys was to test it on a) a more extensive set of images containing munitions, and b) on seabeds in different environments besides the high-coral cover areas used in Shihavuddin et al. (2013). Objectives 1 and 2 did not require altering the existing code, but we could already foresee many ways improvements could be made (Fig. 1 gray box). One way to do so was to include some information about the roughness, or 3D shape, of the seabed in the algorithm.

The topographic structure of a scene can be inferred optically in several ways, including stereo cameras, structured light, or with a single moving camera, using a class of methods referred to as monocular structure from motion (SfM). All three methods have pros and cons and each would be addressed in a full treatment of the problem (Fig 1). In this project, we focused only on structure from motion, however, in order to illustrate the benefits of including topographic relief in the classification process without requiring any additional hardware to acquire the data.

Using SfM, structure was recovered by simultaneously searching for the camera positions and points on the seabed that best complied with the observed images. The microbathymetry (*i.e.* cm-scale bathymetry) recovered from SfM can assist scene understanding and automated classification because relief is independent and complementary to color and texture, which are the sources for 2D features used in classification.

A considerable body of literature exists on the topic of classification from sonar bathymetry (Atallah and Smith 2004) and backscatter (Foster et al. 2013), which are analogs to optical microbathymery and image color/texture. There are very few previous studies using of optical microbathymetry for characterizing natural underwater environments, however. Recent work by Friedman et al. (2012) illustrated the measurement of multiscale rugosity from stereo imaging, but it was not used for feature extraction or automated classification. Objective 3 helped fill this gap by combining texture and structure information in the classification process and applying it to UWMM.

3 Materials and Methods

3.1 Dataset 1: Ordnance Reef, Hawaii

The Ordnance Reef dataset was collected by the National Oceanic and Atmospheric Administration (NOAA) at the "Ordnance Reef" site off of Waianae, Hawaii by divers using a hand-held high-definition video (HDV) camera. Divers swam a series of parallel transects over an area of approximately 20 x 20 m collecting video with the camera pointed vertically down at the seabed. During processing, individual frames were grabbed from the video stream and subsampled by removing every other scan line to remove interlacing effects because the camera, a Sony HDR-XR520V, did not have progressive scan capability. The resulting frames were then composited into a mosaic using the technique described by Lirman et al. (2007). The resolution of the Ordnance Reef mosaic (Fig. 2) was 7,949 × 8,444 pixels at ~2.5 mm per pixel.

3.2 Dataset 2: Sand, Reef, Seagrass Sites, Miami

A team of WRT personnel and UM divers collected underwater images during the last week of May 2014 at four sites off the coast of Key Biscayne, FL (Fig. 3). Unexploded ordnance (UXO) surrogates and inert UXO were used as test objects (Fig. 4). The four sites were chosen based on seabed types. Three sites were used to gather data over different seabed types: bare sediment (sand), seagrass, and coral reef. A fourth site was chosen in an attempt to acquire data along the transitions between these seabed types: seagrass gradually transitioning into sand, for example. We were successful at finding a site with sand, seagrass, and low-relief hardbottom with some corals all within reasonable proximity of one another (~30-40 m radius). The actual conditions at this "mixed" site ended up having very distinct boundaries between the seabed types, however. Therefore, the data from all four sites fell into one of three categories (sand, seagrass, or reef, *e.g.* Fig. 5).

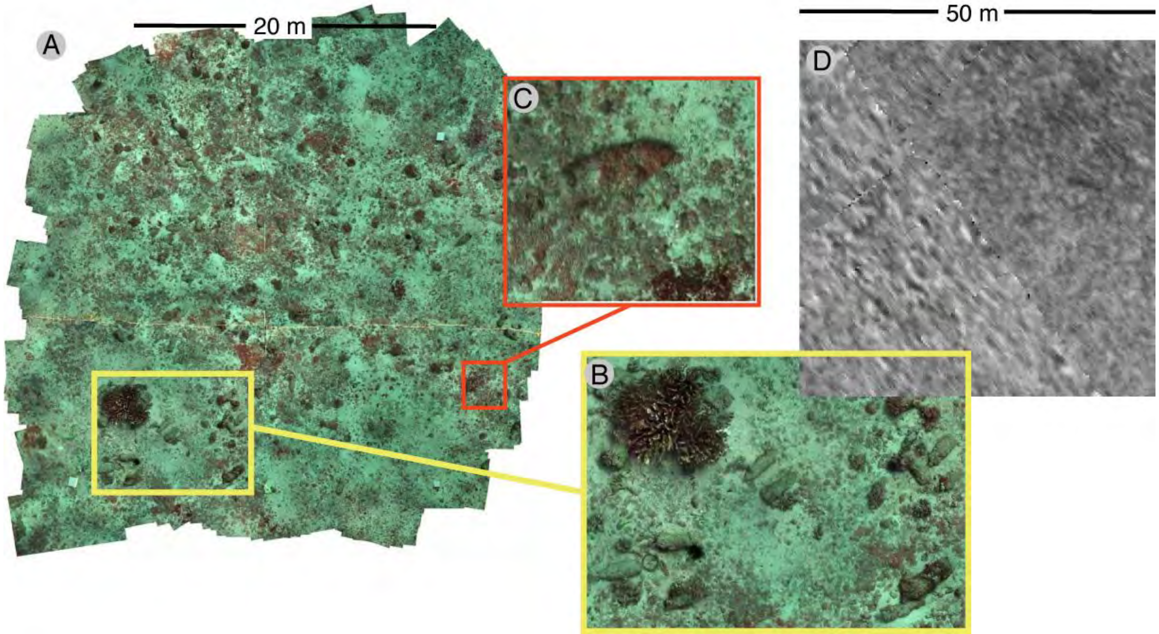


Figure 2. Ordnance Reef mosaic: (A) Landscape mosaic from Ordnance Reef off of Waianae, Hawaii depicting branching, lobate, and encrusting corals, a variety of fish, and several 5-inch projectiles proud on the bottom. Total area covered is approximately 450 m^2 with 2.5 x 2.5 mm pixels. Insets (B) and (C) show portions of the landscape mosaic at full resolution and give an idea of the detail available. The high resolution provided by underwater imagery affords UWMM detection, discrimination from background clutter, and classification. What is needed is a method to reliably automatically segment the optical images. (D) Side-scan sonar image including the same area shown in (A). Total area covered is 2500 m^2 with 30 x 30 cm pixels. Note how much more difficult it is to identify targets due to the lower resolution of the side scan image.

At each of the four sites, the test objects were lowered to the seabed then arranged by divers into two parallel rows containing 8 targets each. The targets were separated by ~2 m along each line and the two lines were themselves separated by ~3 m. The goal was to place the objects in such a way that when taking images from an overhead viewpoint, most of the images contained one target, but some contained none, and some contained two. Once the objects were placed, a diver swam over them with an array of cameras. After the diver with the cameras made a pass over the targets, a second diver then moved all of the targets ~1 m in the same direction, then the diver with the cameras made another pass. In this way, the parallel lines of targets migrated across the seabed and many images of each target were acquired in different poses and over different small patches of seabed. Using this technique we achieved a large sample size of replicate images of targets with only a limited number of known objects.

Data were collected simultaneously using four cameras, enabling the investigation of the effects of resolution on classification accuracy. Two GoPro Hero 3+ cameras were used, with one set to record high-definition (1080 x 1920 pixels), progressive-scan video, and the other set to record 11 Megapixel (3000 x 4000 pixels) still images at 1 s intervals. In addition, two Nikon D7000 digital SLR cameras were used to record 16 Megapixel still images at 1 s intervals; one SLR used a 24 mm lens whereas the other used a 56 mm lens.

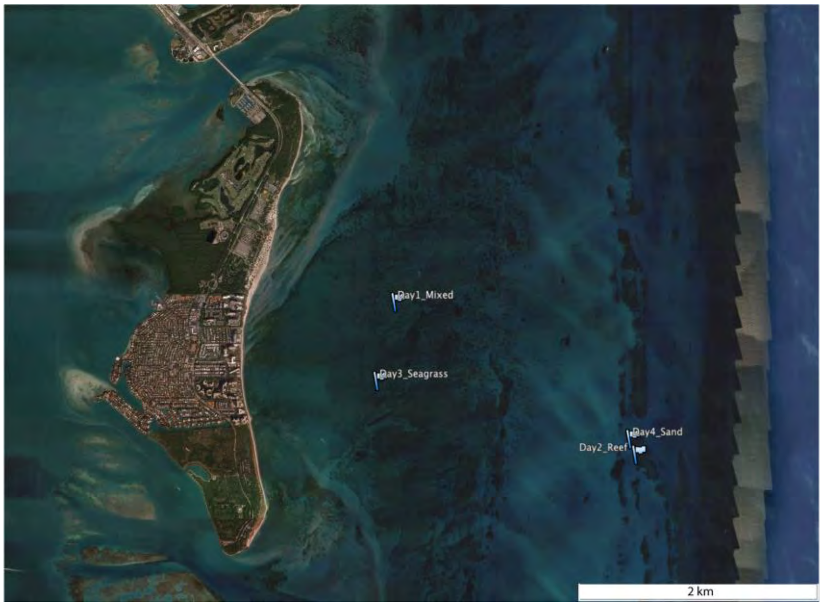


Figure 3. Sites of data collection near Key Biscayne: This image from Google Earth shows the island of Key Biscayne, FL and the locations of the four sites used to collect the Miami dataset.

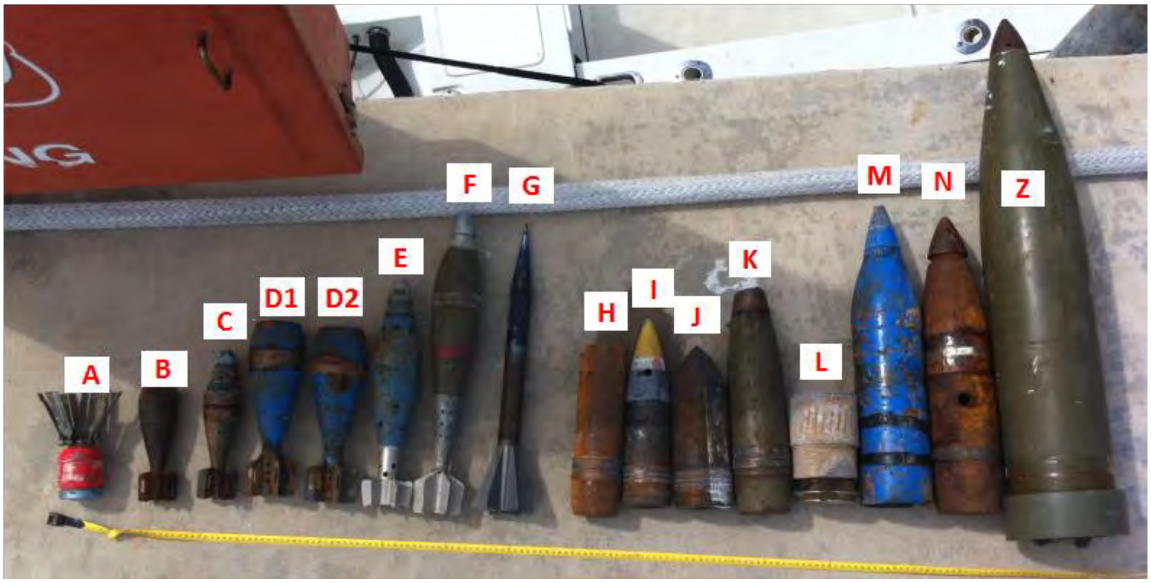


Figure 4. Test targets used for data collection near Key Biscayne: A - BDU-28 Submunition, B - 60mm Mortar (without fuze nose, pitted brown, with tail), C - 60mm Mortar (with fuze, with tail, inert training round; ATC standard target), D1 - 81mm M43A1/M49A2 Mortar (Blue Training Round, No Fuze), D2 - 81mm M43A1/M49A2 Mortar (Blue Training Round, No Fuze), E - 81mm Mortar (with fuze), F - 81mm M821A1/M889A1 (empty) High Explosives (HE) Mortar (with fuze), G - 2" Rocket APFSDS-T M735, H - 76mm Projectile (no fuse), I - 3" Armor Piercing Projectile MK28 Type A, J - 90mm Projectile (no widscreen), K - 90mm High Explosive (HE) Projectile M71, L - APDS Adapter with End Cap and Cartridge, M - 105mm Projectile (Blue Training Round), N - 105mm Projectile (with band and solid tip), Z - 155 mm Howitzer Projectile.

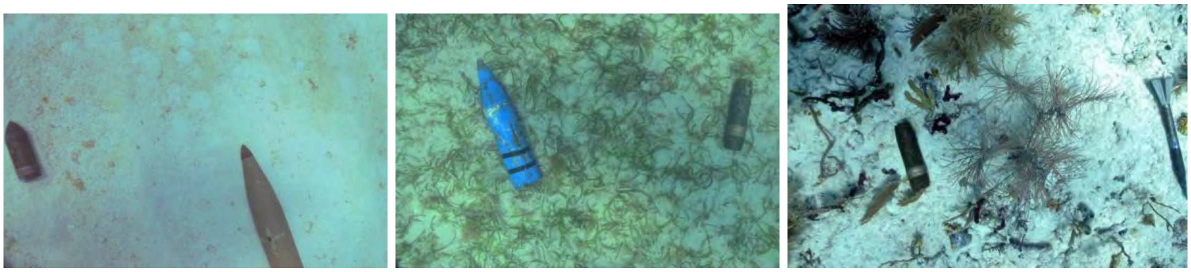


Figure 5. Example images containing test objects on the three seabed types contained in the Miami dataset: Left: 90 mm projectile and 155 mm Howitzer projectile on sand. Center: 105 mm projectile and 76mm projectile on seagrass. Right: 76 mm projectile and 2" rocket on reef.

3.3 Dataset 3: HUMMA (Deep Water), Hawaii

The Hawaii Undersea Military Munitions Assessment (HUMMA) project developed methods and collected extensive datasets designed to evaluate the condition and potential impacts of both conventional and chemical munitions on the underwater environment and the impacts of the underwater environment on the munitions (Edwards et al. 2012). One of the datasets collected during HUMMA consisted of nadir-view images taken by a drop-camera system developed by the Woods Hole Oceanographic Institution (WHOI) and operated by WHOI staff during the HUMMA cruises. This instrument, named the "TowCam", was a deep-water digital camera system towed and connected to a ship via a standard 0.322-inch (8.18-mm) coaxial cable (Fornari and The WHOI Towcam Group 2003). Still images from its 16-megapixel Nikon D7000 DSLR camera were 4928 x 3264 pixels and were acquired at 10 second intervals. Illumination for the photographs was provided by a Benthos 383 strobe electronics unit and two Benthos 386 flash heads with each head providing 300 watt/seconds of illumination. The TowCam had a large "tail" that provided stability and the coaxial cable provided the winch operator with access to real-time depth and altitude information that allowed the system to be steadily towed at an altitude of 4-8 m above the seafloor with each photograph capturing a bottom area of between 15 and 35 square meters. While deployed, the TowCam was tracked using the ship's GPS navigation and an ultra-short baseline acoustic positioning system. The TowCam traversed the area at about 0.5 ± 0.25 knots (0.9 ± 0.45 km/hr). Two green scaling lasers in the camera's field of view provided a 20-cm length reference on each of the photos.

The 17 nighttime surveys conducted during HUMMA with the TowCam from November 23 to December 4, 2012, required 72 hours to complete and provided linear coverage of approximately 109 km. Over 32,000 images were collected, which exceeded a manageable number that could be analyzed for this initial report. Thus, for testing the classification algorithms we used a subset of the images, which had previously been inspected by a human analyst. Kelley et al. (in press) inspected the images from four of the 17 transects (Fig. 6).

The vast majority of the images in HUMMA transects TC-10, TC-13, TC-15, and TC-17 contained only bare sediment. Nevertheless, Kelley et al (in press) identified 221 images with at least one object in addition to the bare substrate. We grouped these objects into one of 7 types following the identifications done by Kelley et al (in press): Natural, M47 bomb, bomb – other, munition - projectile, munition – other, debris – all types, unknown (Fig. 7, Table 1). Then, in order to try to have roughly even numbers of objects, we randomly selected as many as possible up to 12 images containing each type of object to use as a training set and as many as an additional 2 images of each type to use as a validation dataset.

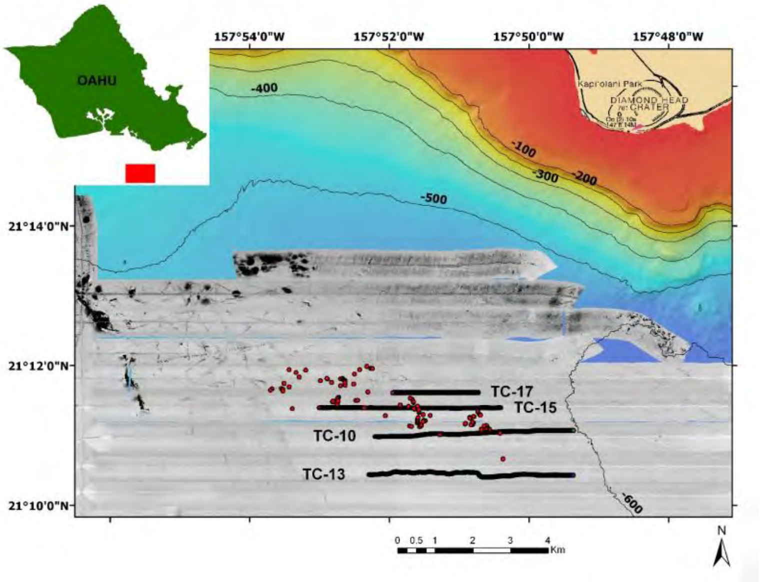


Figure 6. Locations of the four transects that were the source for the HUMMA images used in this study: Mixed multibeam (colored) and side scan (grey scale) map of the study site showing the TowCam transects (TC-XX), locations of known M47A2 bombs (red dots), bathymetric contours in meters (black lines). Image source: Kelley et al. (in press).

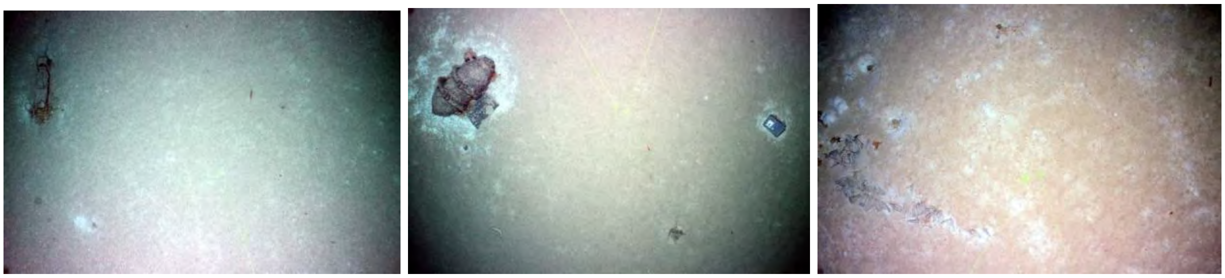


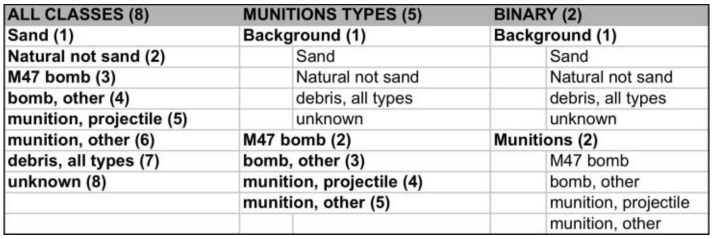
Figure 7. Example images from the HUMMA dataset: These full-frame images from the TowCam cover areas of about 7 x 5 m and have been contrast-stretched to enhance detail. Left: image containing a M47A2 bomb. Center: image containing ordnance classified for our purpose as "bomb-other" in the lower-left and a box classified as "debris" in the center-right. Right: image containing multiple objects classified as "munitions – projectile."

3.4 Image Classifier 1: Original Girona Algorithm

All three of the datasets described above were processed using our existing image classification algorithm (Shihavuddin et al. 2013), hereafter referred to as the "Girona algorithm." The Girona algorithm used a supervised classification approach. Thus, for each dataset, we defined a set of classes and a set of analyst-defined points (*i.e.* "reference" or "ground truth" points). The points picked by the analyst were first randomly selected. Hardly any random points fell on the test targets, however, because as a percentage of the area covered by the dataset the "background" classes dominated. Thus, additional, non-random ground truth points were added by the analyst to ensure that at least 100 points were defined for each class being used. Note, the background classes, for example sand, had perhaps an order of magnitude more points defined in each dataset than the rarest of target classes.

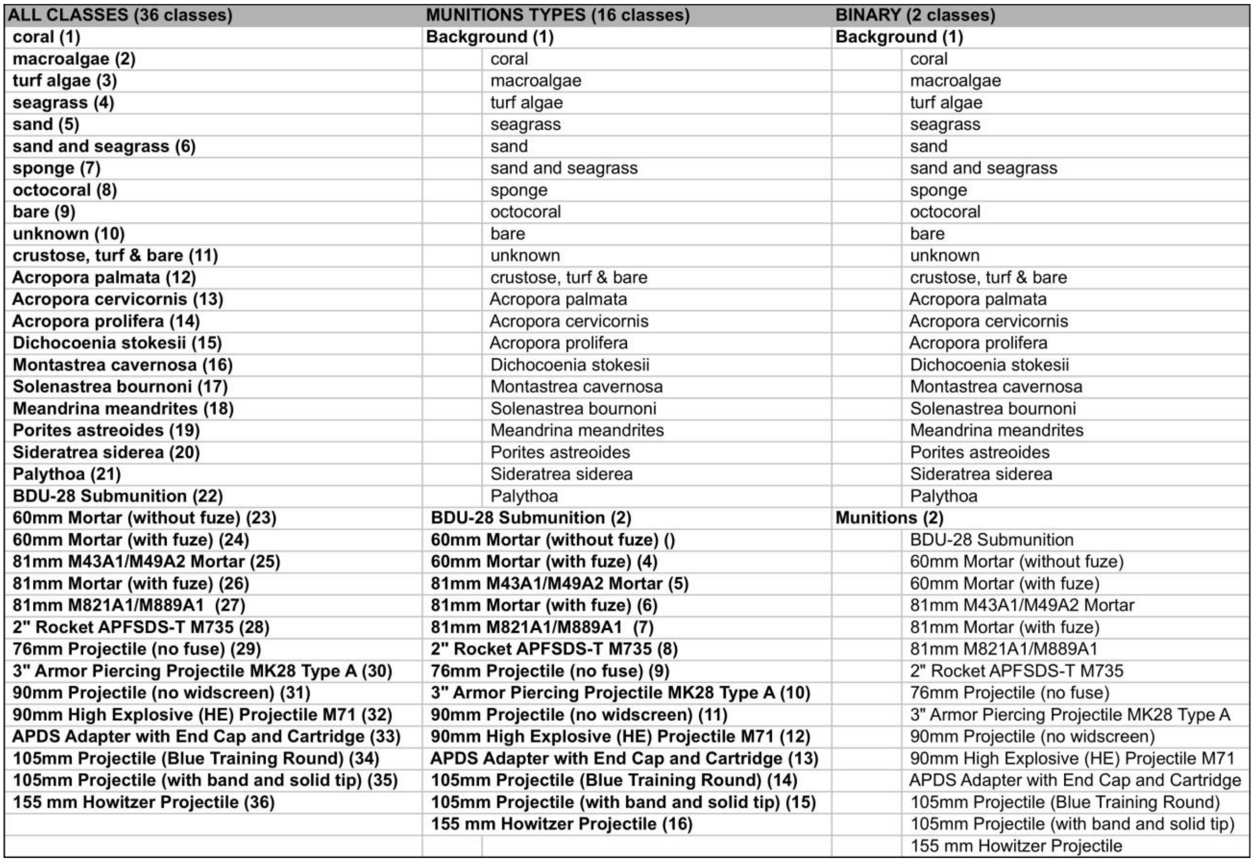
Only two classes were used for the Ordnance Reef mosaic: 5-inch shells and background. The HUMMA dataset used 8 classes (Table 1) and the Miami dataset used 36 classes (Table 2).

Table 1. Classes used for the HUMMA dataset: All 8 classes were used for processing with the Girona algorithm. As part of analyzing the results, these classes were consolidated into Background and Munitions Types (5 classes) and also a binary Munitions / Not Munitions classification scheme.



ALL CLASSES (8)	MUNITIONS TYPES (5)		BINARY (2)	
Sand (1)	Background (1)		Background (1)	
Natural not sand (2)		Sand		Sand
M47 bomb (3)		Natural not sand		Natural not sand
bomb, other (4)		debris, all types		debris, all types
munition, projectile (5)		unknown		unknown
munition, other (6)	M47 bomb (2)		Munitions (2)	
debris, all types (7)	bomb, other (3)			M47 bomb
unknown (8)	munition, projectile (4)			bomb, other
	munition, other (5)			munition, projectile
				munition, other

Table 2. Classes used for the Miami dataset: All 36 classes were used for processing with the Girona algorithm. As part of analyzing the results, these classes were consolidated into Background and Munitions Types (15 classes) and also a binary Munitions / Not Munitions classification scheme.



ALL CLASSES (36 classes)	MUNITIONS TYPES (16 classes)		BINARY (2 classes)	
coral (1)	Background (1)		Background (1)	
macroalgae (2)		coral		coral
turf algae (3)		macroalgae		macroalgae
seagrass (4)		turf algae		turf algae
sand (5)		seagrass		seagrass
sand and seagrass (6)		sand		sand
sponge (7)		sand and seagrass		sand and seagrass
octocoral (8)		sponge		sponge
bare (9)		octocoral		octocoral
unknown (10)		bare		bare
crustose, turf & bare (11)		unknown		unknown
Acropora palmata (12)		crustose, turf & bare		crustose, turf & bare
Acropora cervicornis (13)		Acropora palmata		Acropora palmata
Acropora prolifera (14)		Acropora cervicornis		Acropora cervicornis
Dichocoenia stokesii (15)		Acropora prolifera		Acropora prolifera
Montastrea cavernosa (16)		Dichocoenia stokesii		Dichocoenia stokesii
Solenastrea bournoni (17)		Montastrea cavernosa		Montastrea cavernosa
Meandrina meandrites (18)		Solenastrea bournoni		Solenastrea bournoni
Porites astreoides (19)		Meandrina meandrites		Meandrina meandrites
Sideratrea siderea (20)		Porites astreoides		Porites astreoides
Palythoa (21)		Sideratrea siderea		Sideratrea siderea
BDU-28 Submunition (22)		Palythoa		Palythoa
60mm Mortar (without fuze) (23)	BDU-28 Submunition (2)		Munitions (2)	
60mm Mortar (with fuze) (24)	60mm Mortar (without fuze) ()			BDU-28 Submunition
81mm M43A1/M49A2 Mortar (25)	60mm Mortar (with fuze) (4)			60mm Mortar (without fuze)
81mm Mortar (with fuze) (26)	81mm M43A1/M49A2 Mortar (5)			60mm Mortar (with fuze)
81mm M821A1/M889A1 (27)	81mm Mortar (with fuze) (6)			81mm M43A1/M49A2 Mortar
2" Rocket APFSDS-T M735 (28)	81mm M821A1/M889A1 (7)			81mm Mortar (with fuze)
76mm Projectile (no fuse) (29)	2" Rocket APFSDS-T M735 (8)			81mm M821A1/M889A1
3" Armor Piercing Projectile MK28 Type A (30)	76mm Projectile (no fuse) (9)			2" Rocket APFSDS-T M735
90mm Projectile (no widscreen) (31)	3" Armor Piercing Projectile MK28 Type A (10)			76mm Projectile (no fuse)
90mm High Explosive (HE) Projectile M71 (32)	90mm Projectile (no widscreen) (11)			3" Armor Piercing Projectile MK28 Type A
APDS Adapter with End Cap and Cartridge (33)	90mm High Explosive (HE) Projectile M71 (12)			90mm Projectile (no widscreen)
105mm Projectile (Blue Training Round) (34)	APDS Adapter with End Cap and Cartridge (13)			90mm High Explosive (HE) Projectile M71
105mm Projectile (with band and solid tip) (35)	105mm Projectile (Blue Training Round) (14)			APDS Adapter with End Cap and Cartridge
155 mm Howitzer Projectile (36)	105mm Projectile (with band and solid tip) (15)			105mm Projectile (Blue Training Round)
	155 mm Howitzer Projectile (16)			105mm Projectile (with band and solid tip)
				155 mm Howitzer Projectile

The Girona algorithm classified all the data into one of the 2 (for Ordnance Reef), 8 (for HUMMA), or 36 (for Miami) classes. Results assessment (Section 3.7), was conducted first on the classification using all the classes but then also on reduced sets of classes we called "munitions types", which was all background consolidated into one class, and finally on "binary" classes, which was munitions vs. not munitions (Tables 1, 2).

3.5 Image Classifier 2: WRT Feature Sets

The Girona algorithm (Shihavuddin et al. 2013) was based entirely on image color and texture features. Despite incorporating a wide variety of such features, one question regarding such an approach was always whether some other set of feature descriptors would work better. In order to address this question, we assessed the potential of a different set of features, which were derived by the WRT team based on previous success in radar and IR image analyses (Kreithen et al. 1993; Dell'Acqua and Gamba 2003; Ho et al. 2008; Nakatsu et al. 2012).

The image-based features developed by WRT were derived from several categories: physics-based, statistical-based, correlation-based, texture-based, wavelet and fractal-based, and energy-based. Each category provided multiple methods for measuring similarities and differences between targets and clutter. These features were distinct from the physics-based radar features that measure physical properties of the targets and were extracted from phenomenological models. Statistical-based features measured the statistical attributes of a region such as mean value, standard deviation, skewness, and kurtosis. Correlation-based features measured the level of spatial correspondence between target and clutter objects. Texture-based features provided information regarding the spatial interrelationships and arrangement of basic elements and were based on image histogram, gray-level co-occurrence matrix, gradient matrices, and wavelet transforms. Fractal-based features were used to estimate fractal and non-fractal behavior. Energy-based features utilized intensity-, amplitude-, or texture-based statistics to characterize detected regions. For each of these categories, multiple features were computed. In total, 63 image features were calculated, each from one of the three categories below:

3.5.1 Texture Features via Gray-Level Histograms, Co-occurrence, and Run Length: In first-order statistical feature analysis, information on texture was extracted from the histogram of image intensity values. This approach measured the frequency of a particular gray-level but did not account for spatial correlations, or co-occurrences, between pixels. Its main advantage was simplicity; using N distinct gray levels to compute moments of a pixel occurrence probability was conceptually and computationally simple. Seven features describing the properties of the gray level histogram were computed: mean, variance, coarseness, skewness, kurtosis, energy, and entropy.

Second-order statistic texture analysis yielded information on 2-D texture based on the probability of finding pairs of gray levels at set distances and orientations over a region of interest image. Properties of the gray-level co-occurrence matrix (GLCM) have been extensively evaluated for deriving meaningful features from radar and electro-optical features (e.g. Haralick et al. 1973; Shanmugan et al. 1981; Ulaby et al. 1986; Dell'Acqua and Gamba 2003). Our basic feature computations for GLCM textures were based on Haralick's 14 local features and extension thereof including contrast, mean correlated energy, entropy, homogeneity, cluster prominence, cluster shade, symmetry, angular second moment, smoothness, dispersion, among others. We also implemented a similar but distinct approach following the texture feature coding method (TFCM) proposed by Horng et al. (2002). TFCM combines information from the gray level histogram and GLCM and translates an intensity image to a texture feature number matrix using gradient images and center-symmetric auto-correlations. Using this new type of co-occurrence matrix, we computed a number of features such as: coarseness, homogeneity, mean convergence, code entropy, and code similarity.

Finally, we also extracted some features using extensions to higher-order statistic, but limited our approach based on the tradeoff between information value and increasing complexity with the number of variables investigated. The gray level run length method was based on higher-order statistics, where information on a gray level is developed over a particular vector "run" through the image such as an anomaly boundary. Coarse textures were derived from long runs and fine textures are dominated by short runs. We computed a number of short and long run features for varying run paths and directions: short run emphasis, long run emphasis, run length uniformity, and run length percentage.

3.5.2 Spatial Interest Pixels and Signal-to-Clutter Ratio Features: In addition to the low-level texture features, we also assessed the potential of spatial interest pixels (SIPs) or interest points (Li et al. 2003). A SIP is a pixel of region or points that constitute a strong interest strength relative other pixels or groups of pixels. We utilized a version of interest point computation based on the Harris detector (Schmid et al. 2000), which derived the interest strength from the eigenvalues of the gradient correction matrix. The interest strength of a local window or mask was determined and used to compute statistical measures of the SIP: average interest strength, maximum interest strength, and strength variance.

3.5.3 Fourier Descriptors, Wavelet, and Fractal-based Image Features: Fourier analysis and wavelet or fractal decomposition can be used to extract features related to the texture, spatial wavelength, self-similarity and fractal dimension of the radar return or region of interest signal anomalies. The fractional Fourier transform, in particular, was used to map the time-frequency distribution of images into fractional domains. Our algorithmic implementation utilized short-time Fourier transforms and Hu moments to compute Fourier descriptors and invariant moments as shape features. In addition fractal geometry (Turcotte 1997) was used to extract features associated with self-similarity and fractal scale. Specifically, we calculated the box-counting dimension and related Huasdorff dimension of each region of interest.

3.6 2.5-D Extensions to the Girona Image Classifier

As mentioned above, the Girona algorithm (Shihavuddin et al. 2013) was based entirely on image color and texture features. Many other types of features could be added to extend this basic approach, for example: shape, symmetry, absolute scale, and contextual information, among others. The purpose of experimenting with 2.5-D extensions to the Girona image classifier was to add features to the algorithm based on the height of the seabed. The idea was to see if adding extra information helped improve the performance of the basic Girona algorithm. The fundamental difference between this effort and the testing of additional features described above (Section 3.5) was that addition of height is a fundamentally new type of data whereas the additional 2-D features (Section 3.5) were new ways to process the same essential source data.

A note on the terminology; we call the digital elevation model (DEM) approach used here a “two and a half dimensional” (2.5-D) method because it only captures information that can be seen from the overhead viewpoint of the camera. In contrast, a true three-dimensional (3-D) method would require data from all viewpoints around an object and would therefore capture information from vertical or overhanging surfaces that can't be seen from overhead. The rest of this section outlines the 2.5-D DEM generation and features. For details, see Shihavuddin et al. (2014), included as an appendix to this report.

The first step to incorporating 2.5-D features was to generate a digital elevation map from the original dataset. We used a structure from motion (SfM) technique (see Shihavuddin et al. 2014 for details) to generate the DEM. Note that the SfM technique requires multiple overlapping images for every part of the DEM to be generated. The Ordnance Reef and Miami datasets had sufficiently overlapping images for this purpose, but the HUMMA images were intentionally collected with minimal (~10%) overlap, so were not suitable for this method. Figure 8 shows an example of 2.5-D reconstruction of a portion of the Ordnance dataset.

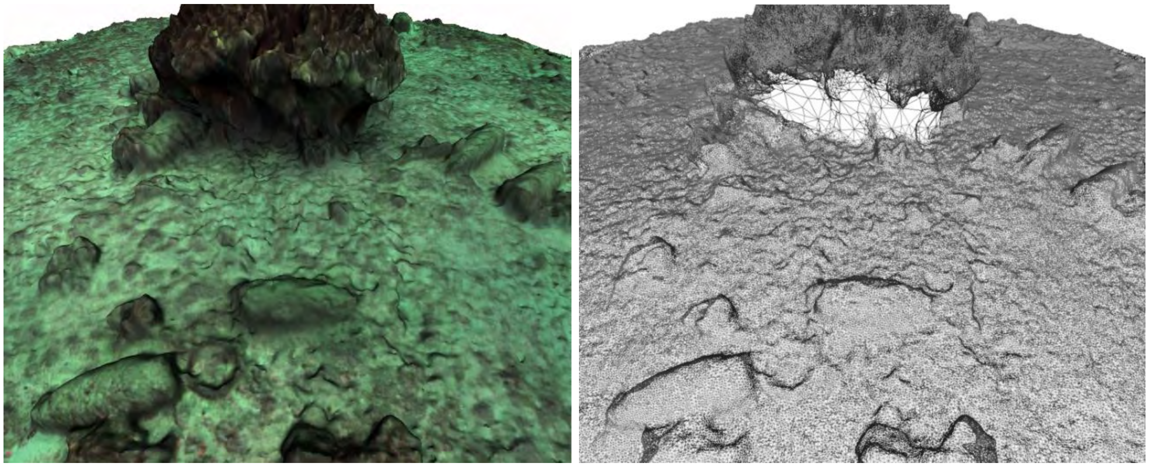


Figure 8. 2.5-D reconstruction of a portion of the Ordnance Reef dataset displayed as a textured surface (left) and as a triangular mesh (right): Scale varies in this oblique view with the lower (closer) edge of the image corresponding to ~2 m on the seabed.

Our method used a patch-based classification framework, where a sliding window was used to sample "patches" around regularly spaced center pixels and classify them based on available 2-D and 2.5-D features. The size of the patches was dictated by the sizes of the important object classes. Generally, the size of the patch should be as small as possible while still being large enough to encapsulate distinct features of an object. At this point, choosing a patch size is still subjective; we tested a range of values and picked the one that worked best. For the Ordnance Reef dataset, 128-pixel square windows produced the best results. We found that the Miami datasets used a slightly larger range of 198 – 300 pixels square to produce the best results. The best window size for the HUMMA data also fell in this range, at 256 pixels square.

The 2.5-D features were polynomial surface coefficients (Keren and Gotsman 1999), elevation statistics (Hetzel et al. 2001), slope of surface (Friedman et al. 2010), curvature (Wang et al. 2000), surface normal (Hetzel et al. 2001), rugosity (Friedman et al. 2010) and symmetry (Kovesi 1997) extracted from the elevation map. Details of each feature are explained in the Appendix (Shihavuddin et al. 2014).

For the Ordnance Reef dataset we used all of the available images to generate a single mosaic image and DEM. The Miami datasets were not collected in a way that was amenable to this approach. Specifically, we were unable to make a single mosaic of all the images at a given site because we moved the munitions after one pass of the cameras. Therefore, for the Miami dataset we picked two "strips" of about 50 images from each of the reef and seagrass datasets for processing. Thus there were four datasets from Miami that were processed with both 2-D and 2-5D; we called these "reef 1", "reef 2", "seagrass 1", and "seagrass 2."

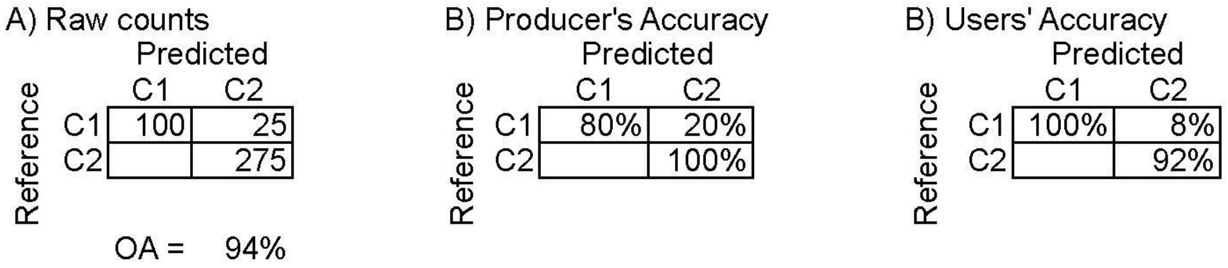
3.7 Classification Assessment

Classification assessment was conducted using the error matrix approach (Congalton and Green 1999). The error matrix uses N points within the classified dataset that have known, verified class identification for comparison with the classification results. The known points are variously called "reference" or "ground truth" data in the remote sensing literature and, in general, they may be used either for training the classifier, testing the classifier, or both.

Ideally, a large sample N_1 would be used for training, and an independent, equally large sample N_2 would be used for evaluating the results. Most often, however, generating the reference data turns out to be one of, if not the most, expensive item in a remote sensing study, so the training sample N_1 turns out to be larger than the evaluation sample N_2.

In this study, the sample sizes N_1 and N_2 varied for the different datasets. The Ordnance Reef dataset had $N = 10{,}681$ and used 10-fold cross validation for accuracy assessment. In 10-fold cross validation, 90% of the N samples were used for training ($N_1 = 9{,}613$), then the remaining 10% were used for validation ($N_2 = 1{,}068$). This was repeated 10 times, generating new training and validation sets with random subsets each time and the results averaged. For the Miami dataset, N_1 and N_2 varied by site and camera because a minimum of N = 400 points were randomly selected for each site and camera combination then a variable number of non-random points were added by the analyst to fill in the less-frequently observed classes. For the Miami dataset $1216 <= N_1 <= 2716$ and $248 <= N_2 <= 548$. Assessment of the Miami dataset did not use cross validation. Instead, the error matrix for each site/camera combination was computed both from the training data and from the validation data. For the WRT features computed from the Miami dataset, $N = 436$ all used for training. Validation for the WRT feature dataset was performed on the training data. For the HUMMA dataset $N_1 = 9818$ and $N_2 = 728$.

A simple example error matrix will help clarify the terminology used in the results, below. Assume a hypothetical dataset had $N = 400$ ground truth points divided between two classes (C1 and C2) as shown in Figure 9A. Normalizing the raw error matrix in different ways produces different measures of the quality of the classification. The simplest metric is the overall accuracy (OA), which is the number of reference points correctly classified divided by the total number of reference points. Normalizing the error matrix by the row or column sums reveals two other metrics called "Producer's" and "User's" accuracy, respectively (Congalton and Green 1999; Figs. 9B, 9C). Normalizing by the row sums, (*i.e.* the number of points in each class in the reference dataset), gives an indication of the rate of *false negatives* on a per-class basis. In our example, we see that class 1 had an 80% accuracy with respect to false negatives; in other words, 20% of the points that should have been class 1 were actually classified as class 2. On the other hand, class 2 had no false negatives; all of the points that should have been classified as class 2, were in fact classified that way. Conversely, normalizing by the column sums, (*i.e.* the number of points in each class in the classified, or predicted, dataset), gives an indication of the rate of *false positives* on a per-class basis. In our example, we see that class 1 was 100% accurate with respect to false positives; every pixel that was classified as class 1 was in fact class 1 in the reference dataset. On the other hand, class 2 was only 92% accurate with respect to false positives; 8% of the points that were classified as class 2 were actually class 1.



A) Raw counts

Reference	Predicted C1	Predicted C2
C1	100	25
C2		275

OA = 94%

B) Producer's Accuracy

Reference	Predicted C1	Predicted C2
C1	80%	20%
C2		100%

B) Users' Accuracy

Reference	Predicted C1	Predicted C2
C1	100%	8%
C2		92%

Figure 9. Hypothetical error matrix for two classes with *N*=400 reference data points: (A) raw error matrix. (B) Producer's accuracy, which is the raw matrix normalized by the row sums. (C) User's accuracy, which is the raw matrix normalized by the column sums. The overall accuracy (OA) is the sum of the raw counts on the diagonal normalized by *N*. Note that producer's accuracy is an indication of false negatives whereas the user's accuracy is an indication of false positives.

The assessment of the WRT feature set involved some extra steps beyond those for the Girona algorithm and its 2.5-D extension. These steps involved dimensionality reduction by assessment of which features were most useful for class discrimination. In fact, these same steps were included in the Girona algorithm, just that they are performed in a non-interactive way using either principal components analysis (PCA) or the Fisher kernel (Shihavuddin et al. 2013).

For the WRT feature set, a combination of manual outlining and semi-automated image segmentation led to identification of 374 regions of interest (ROIs) from the Miami dataset.Of these, 311 ROIs were around background features, and 63 ROIs were around targets. The 63 WRT features (Section 3.5) were computed for each of the 374 ROIs.

The first step of the assessment was to determine which of the 63 features best discriminated between targets and background. In other words, which of the features were useful for classifying munitions vs. background. Non-parametric statistical measures were used to assess the 63 features based on single-feature distribution analyses. The distance measures included Mann-Whitney and Wilcoxon Rank Sum statistics metrics, Information Gain, Mahalanobis and Fisher linear statistics, and the Gini Index.

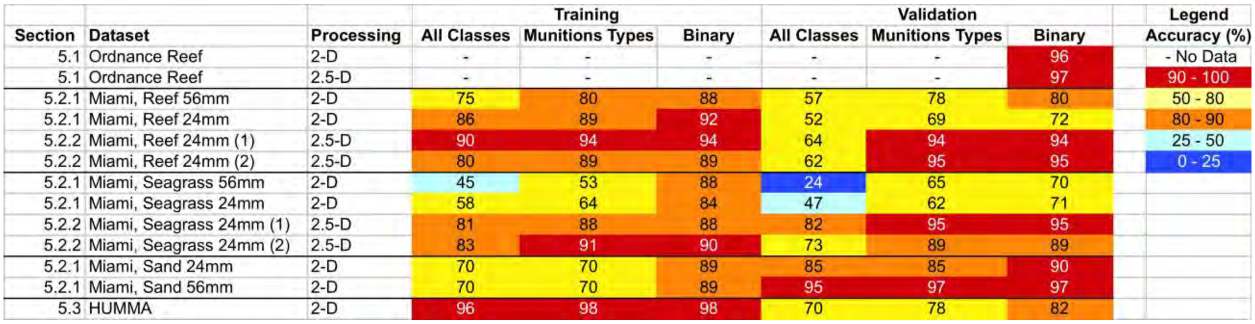
After determining the most useful features, several of the most-promising ones were used to classify munitions vs. background using a discriminative statistical approach. With discriminative analysis, we were not seeking to explain the underlying distributions of the features in a physically meaningful way, but rather we were only concerned with identifying decision boundaries which provided an optimal separation of classes (the classes in this case being munitions and not-munitions). Optimality was defined by minimizing the probability of misclassifying a new feature vector. Examples of discriminative classifiers include Support Vector Machines and Linear Discriminant Analysis. In this work we used a quadratic classifier to discriminate between the classes. Finally, the results of the class separation were assessed using the overall accuracy computed from error matrices, as described above.

4 Results and Discussion

General trends in the dataset were summarized using the overall accuracy (Table 3), which was defined as the number of correctly classified reference points as a percentage of the total number of reference points available for a given dataset (Fig. 9A). One apparent pattern was that OA computed using the training data was higher than OA computed using independent validation data. The exception to this trend was for the Miami data over sand, but later, in Section 4.2.2, we show that although the accuracy of the sand class increased in this case, the accuracies for the individual munitions classes did not all increase over sand using the validation data.

A second clear pattern that emerged from the OA data (Table 3) was that accuracy increased as fewer classes were used (*i.e.* OA of All Classes < OA of Munitions Types < OA of Binary). As shown in Sections 4.2 and 4.3, this general trend was robust at the individual class level also.

Finally, a third trend observed in the OA results (Table 3) was that accuracy for datasets using the 2.5-D algorithm was greater than using just the 2-D algorithm. There was an increase of approximately 20 percentage points in overall accuracy for the two subsets of the Miami reef 24 mm data that were processed with the 2.5-D algorithm relative to the versions processed with just the 2-D algorithm. The increases were closer to 30 percentage points for the two subsets of the Miami seagrass 24 mm data processed with the 2.5-D algorithm relative to the data processed with only the 2-D algorhtm. There was a small increase noted for the Ordnance Reef dataset as well, though less dramatic than the Miami data because the 2-D results were high to begin with at Ordnance Reef.



			Training			Validation			Legend
Section	Dataset	Processing	All Classes	Munitions Types	Binary	All Classes	Munitions Types	Binary	Accuracy (%)
5.1	Ordnance Reef	2-D	-	-	-	-	-	96	- No Data
5.1	Ordnance Reef	2.5-D	-	-	-	-	-	97	90 - 100
5.2.1	Miami, Reef 56mm	2-D	75	80	88	57	78	80	50 - 80
5.2.1	Miami, Reef 24mm	2-D	86	89	92	52	69	72	80 - 90
5.2.2	Miami, Reef 24mm (1)	2.5-D	90	94	94	64	94	94	25 - 50
5.2.2	Miami, Reef 24mm (2)	2.5-D	80	89	89	62	95	95	0 - 25
5.2.1	Miami, Seagrass 56mm	2-D	45	53	88	24	65	70	
5.2.1	Miami, Seagrass 24mm	2-D	58	64	84	47	62	71	
5.2.2	Miami, Seagrass 24mm (1)	2.5-D	81	88	88	82	95	95	
5.2.2	Miami, Seagrass 24mm (2)	2.5-D	83	91	90	73	89	89	
5.2.1	Miami, Sand 24mm	2-D	70	70	89	85	85	90	
5.2.1	Miami, Sand 56mm	2-D	70	70	89	95	97	97	
5.3	HUMMA	2-D	96	98	98	70	78	82	

Table 3. Summary of overall accuracy computed for various datasets and processing options: The overall accuracy (OA) is the percentage of correctly classified points out of the total number of reference points. Note general trends: Training OA was higher than validation OA, except for the Miami sand data. OA increased as fewer classes were used (*i.e.* All Classes < Munitions Types < Binary). OA on datasets using the 2.5-D algorithm was greater than using just the 2-D algorithm.

Even though the overall accuracy (Table 3) gave a good summary of the main results, as discussed in Section 3.7, overall accuracy does not present the full picture of the results because it does not show class-by-class information and because it is skewed if the reference data are not equally distributed among the classes. Therefore, more details for each dataset are presented in the following sections. Results using the original Girona algorithm and the 2.5-D extension to the original Girona algorithm are presented for the first two datasets: Ordnance Reef, Hawaii and the Miami environments. The original Giorna algorithm was also applied to the HUMMA dataset but the 2.5-D extension could not be used with the HUMMA data because there was not enough overlap between successive frames to enable stereo reconstruction. Results for the Miami dataset are also presented using the alternate 2-D image classifier constructed by WRT.

4.1 Ordnance Reef, Hawaii

Using the $N = 10{,}681$ reference points with 10-fold cross validation resulted in an overall accuracy of 96% using the 2-D Girona algorithm and 97% using both the 2-D and 2.5-D features (Shihavuddin et al. 2014). Thus, using the 2.5-D features did improve results, but only by a small amount. In large part, the reason the improvement was small was because the results with 2-D were so good to begin with. Overall accuracy approaching 100% is extremely high in general, and, in this case, should be interpreted with the understanding that accuracy assessment can be strongly affected by random chance for datasets with small numbers of classes (2 in this case) and / or one dominant class (background was 93% of the total number of pixels). Shihavuddin et al. (2014) accounted for chance agreement in the accuracy assessment with additional statistical measures called the kappa coefficient (Congalton and Green 1999) and the average mutual information index (AMI, Finn 1993), both of which accounted for random agreement when computing accuracy. Both kappa and AMI showed improvements in accuracy using the 2.5-D data over the results using just the 2-D data alone (Shihavuddin et al. 2014).

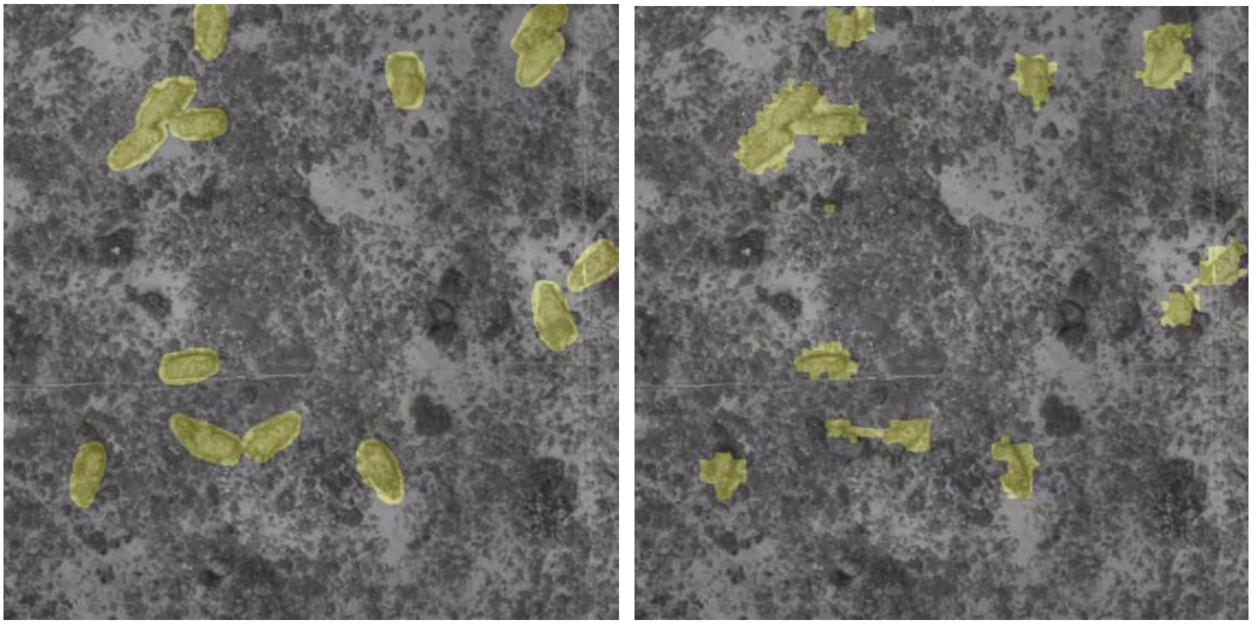


Figure 10. Classified versions of the Ordnance Reef dataset: Background on both the left and right panels shows a portion of the Ordnance Reef mosaic (Fig. 2) in grayscale. Left: munitions identified by an analyst shown in yellow with the borders defined by hand-drawn regions. Right: munitions identified by the 2.5-D extension to the original Girona algorithm shown in yellow. Note edges of the regions defined by the automated segmentation were slightly different than the hand-drawn result, but the number and spatial pattern of the patches agreed well with the hand-drawn result.

4.2 Miami

The Miami dataset was the most complex of the three analyzed in this project, covering multiple habitat types with multiple cameras and including analysis in both 2-D and 2.5-D. In contrast, each of the Hawaii datasets had only one camera and one habitat type. In order to more easily interpret results from the various permutations of habitat and camera in the Miami data, first consider in detail the error matrix for 2-D analysis of one camera in one habitat (Section 4.2.1). This will serve as a reference when interpreting the more complex, consolidated views of multiple error matrices for the 2-D (Section 4.2.2) and 2.5-D (Section 4.2.3) results in the following sections.

4.2.1 Example Calculations from the Miami Reef site: The Nikon D7000 images from the Miami reef site employed a total $N_1 = 2112$ points to train the Girona algorithm. Following classification, 1824 of these training points fell along the main diagonal in the error matrix, resulting in an overall accuracy OA = 86% for the training dataset (Fig. 11). As was typical for the Miami data, this error matrix had many classes with no data (gray elements of Fig. 11). The primary reason for cells with no data was that not all habitat types were present in each dataset. For example, rows 2-4 were empty in the raw counts for this error matrix because these classes (macroalgae, turf algae, and seagrass, see Table 2) were not present in the reef dataset. Furthermore, some classes were subdivided more finely than necessary, for example classes 12-20 were various species of corals (see Table 2), not all of which were found at this reef. Note that most of the munitions were present, except for class #26, 81mm mortar (with fuze). Recall that the images used for each test were randomly selected from the full dataset, so even though the 81mm mortar with fuze was present at the reef site, it just happened to not be in the subset of images that we processed.

As discussed and shown above (Fig. 9), the raw error matrix can be normalized by either row or column sums to generate “producer’s” or “user’s” accuracy, respectively. After doing both, high values along the main diagonals indicated that most classes had both producer’s and user’s accuracy > 80% for the training data used at the Miami reef site, Nikon D7000 with 24 mm lens (Fig. 12). The munitions component of the error matrix, in particular, had high accuracy values (lower right quadrant of Fig. 12). Almost all of the munitions - not just munitions as a whole, but the individual types - were classified with accuracy > 80%.

Interpretation of the producer’s and user’s accuracies together gives the best picture of classification performance (Congalton and Green 1999). For example, the coral class (class #1; Table 2) had relatively low producer’s accuracy, about 25%, and very high user’s accuracy, 100% in fact (Fig. 12). Low producer’s accuracy indicated that many pixels that should have been classified as coral were actually classified as something else, octocoral (class #8) and crustose, turf & bare (class #11), in this case. High user’s accuracy indicated that all of the pixels classified as coral were actually coral. Taken together this indicated that the coral class was underclassified. In other words, the coral class had many false negatives but no false positives.

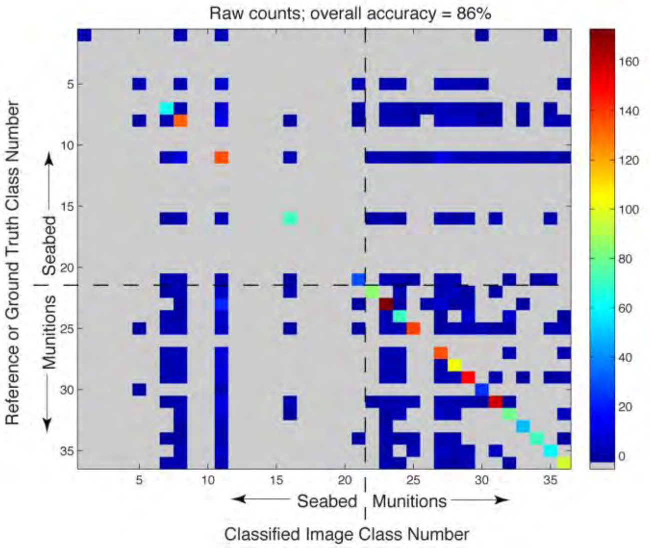


Figure 11. Error matrix of *N* = 2112 raw counts for the Miami dataset, reef site, Nikon D7000 with 24 mm lens: Rows and columns are numbered 1-36 corresponding to the class definitions given in Table 2 (all classes). Dashed lines indicate the boundaries between habitat (#1-21) and munitions (#22-36) classes. The color of each cell indicates the number (see the colorbar) of sampled points that were defined in each class by an analyst (rows) and by the Girona algorithm (column). Gray cells indicate there were no sampled points with the given ground truth / classified data combination. Note that 86% of the sampled points fell along the main diagonal, which means the algorithm-derived class generally agreed with the ground truth.

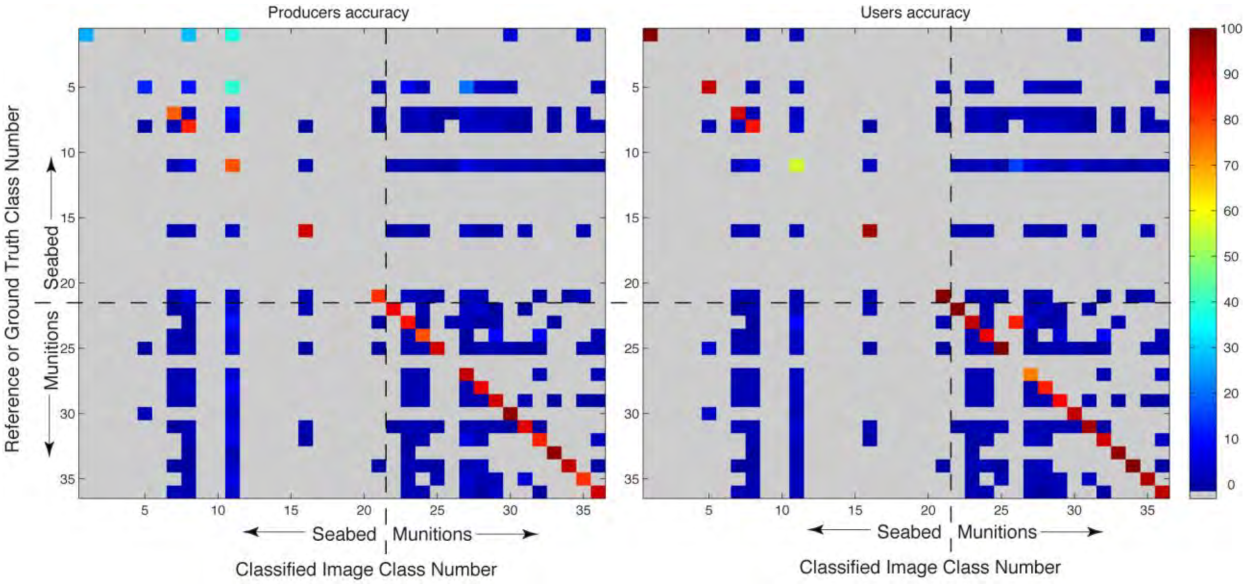


Figure 12. Normalized error matrix for the Miami dataset, reef site, Nikon D7000 with 24 mm lens: Left: Producer's accuracy, which is the error matrix from Fig. 11 normalized by row sums. Right: user's accuracy, which is the error matrix from Fig. 11 normalized by column sums. Colorbar indicates accuracy as a percentage. Note the red elements along the main diagonal indicate that most classes had both producer's and user's accuracy > 80%.

Reducing the number of classes is equivalent to consolidating elements of the error matrix. For example, the matrix shown in Figures 11 and 12 evaluated all 36 habitat and munitions classes considered in the Miami dataset (Table 2). However, if we were only interested in munitions types and willing to place everything else into a "background" class, we could consolidate rows and columns 1-21. Likewise, if we were only interested in a binary munitions / not munitions class scheme we would consolidate rows and columns 1-21 and rows and columns 22-36. When the Miami 24 mm reef dataset was consolidated in this way, the accuracies of the consolidated classes increased (Fig. 13). It is generally the case that using fewer classes results in higher accuracy.

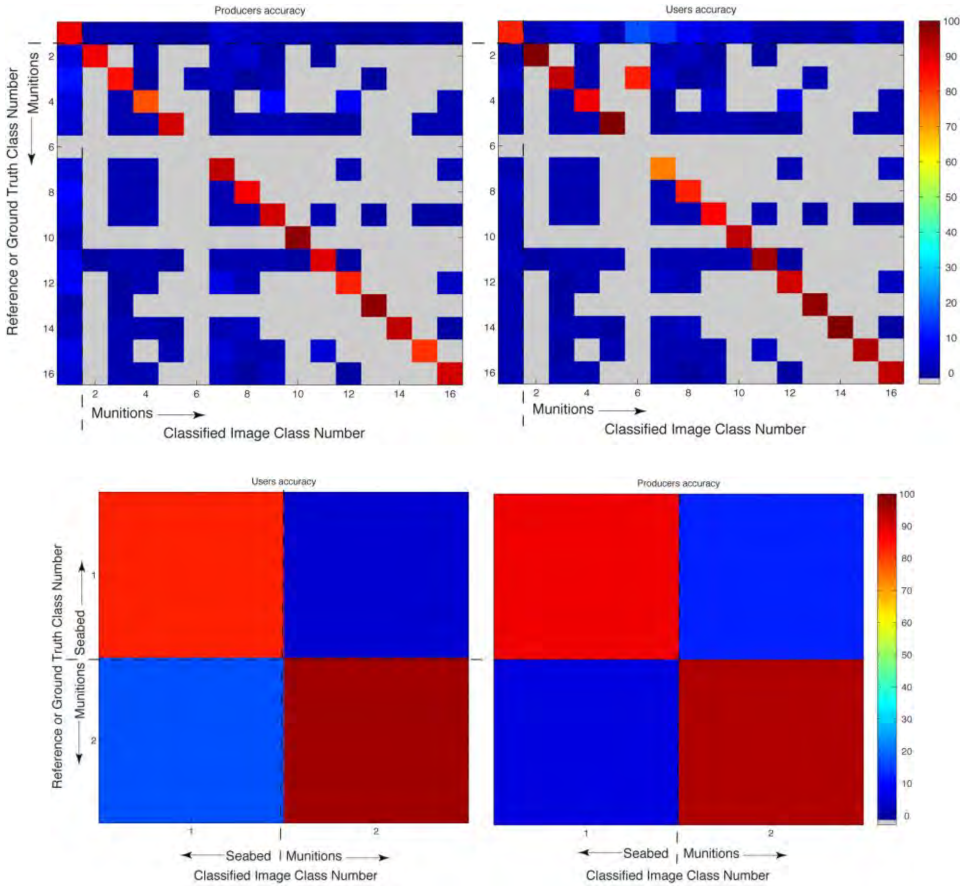


Figure 13. Two versions of a consolidated, normalized error matrix for the Miami dataset, reef site, Nikon D7000 with 24 mm lens: All four of these were derived from the raw counts in Fig. 11. Upper row: Producer's (left) and user's (right) accuracy reduced to background and munitions types classes (Table 2). Lower row: Producer's (left) and user's (right) accuracy reduced to binary presence/absence classes (Table 2). Colorbar indicates accuracy as a percentage. Note that accuracy increased relative to the version with all classes (Fig 12).

4.2.2 Miami Dataset 2-D Results: The complete, normalized error matrix illustrated not only which classes have low accuracy, but also the nature of the misclassification (Figs. 12, 13). Most of the information in the error matrix can be gathered just from the main diagonal, however. Displaying the main diagonals of multiple error matrices side-by-side provided a convenient way to compare the performance of multiple datasets (Fig. 14). Each column in figure 14 was extracted from an error matrix for a different camera / site combination, which are labeled along the bottom axis. Each row in figure 14 corresponds to a class defined in Table 2. To help navigate the plot, dashed lines have been added to demark the seabed classes (rows #1-21) and the munitions classes (rows #22-36). Dashed lines also subdivide the plots by site (reef, seagrass, sand). The two columns within each site presented the results from the two D7000 cameras.

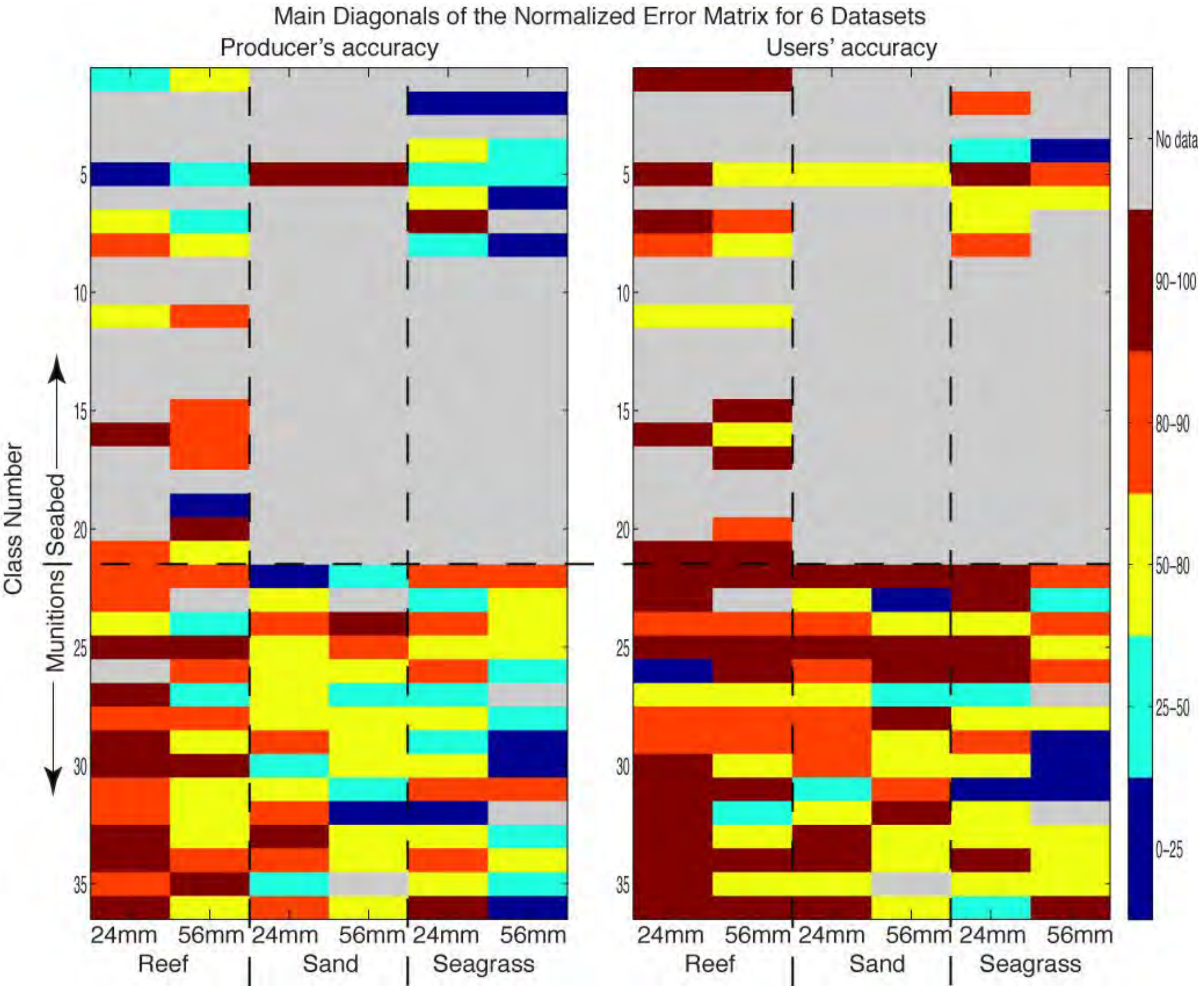


Figure 14. Main diagonals of the normalized error matrix assessed with the data used for training for two cameras at all three of the Miami dataset sites: The first column in both the left and right panel of this figure contains the cells from the main diagonals shown in Figure 12. The other columns are from similar error matrices computed for the camera (D7000 with 24 or 56 mm lens) and site (reef, sand, seagrass) combinations shown. Colorbar indicates accuracy as a percentage. Rows represent class numbers that were defined in Table 2.

Overall, the results of the accuracy assessment with the data used to train the classifier were promising (Fig. 14). The median producer's and user's accuracies for all cameras and sites were 83% and 71% respectively, which was very high for 36 classes.

Results for the identical analysis using an independent validation dataset (Fig. 15) did not have as high accuracy as those using the training dataset (Fig. 14). Note that some of the classes were highly accurate even in the validation dataset. For example, sand (class #5) at the sand site (middle pair of columns) was nearly perfect (Fig. 15). Sand, and mixed sand and seagrass (classes #5, 6) were also highly accurate in the seagrass site (Fig. 15). Some of the munitions had high accuracy, whereas others had very low accuracy (Fig. 15).

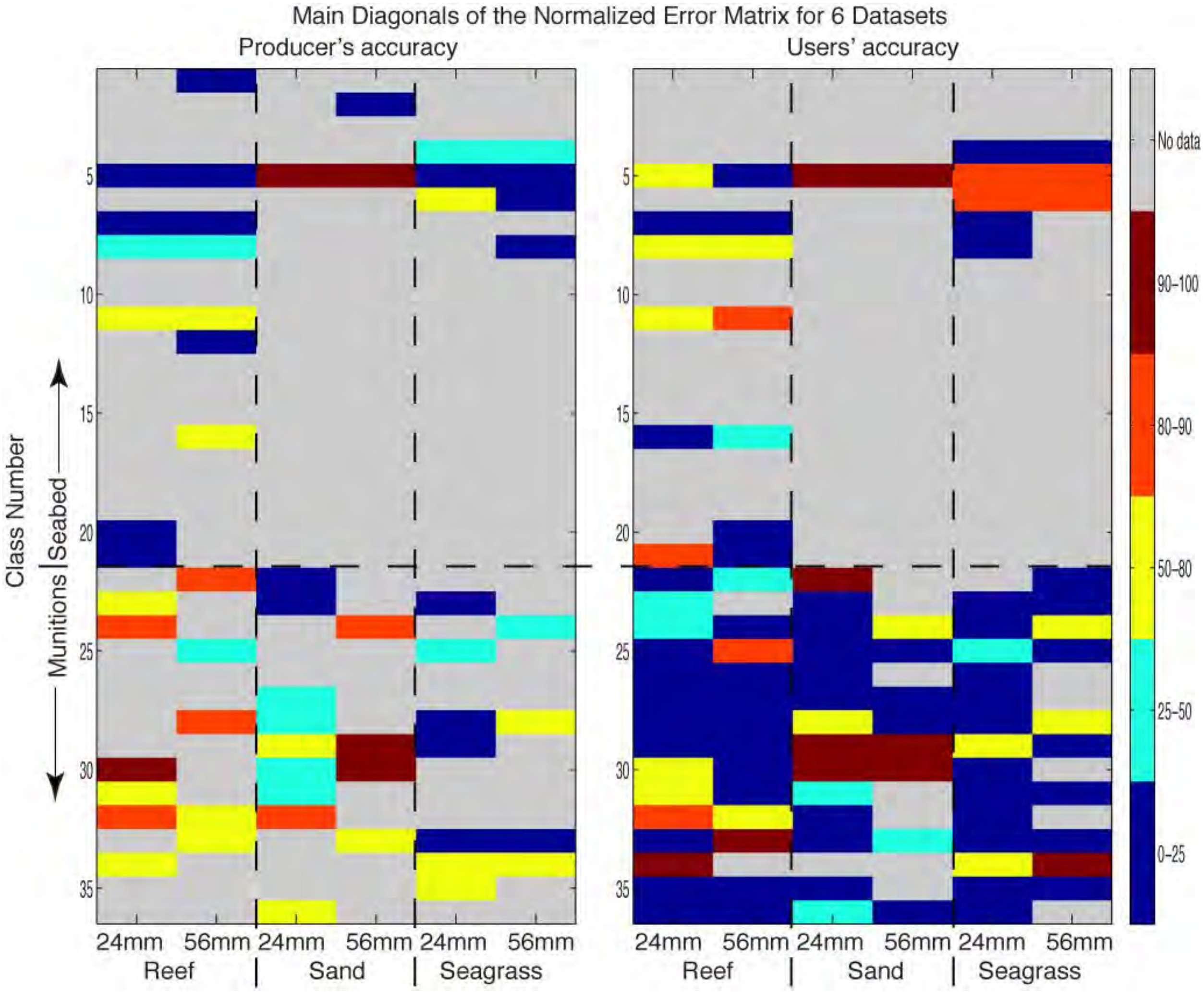


Figure 15. Main diagonals of the normalized error matrix assessed with independent validation data for two cameras at all three of the Miami dataset sites: This figure is the same as Figure 14 except that the data used for validation were independent of the data used for training the classifier.

Consolidating the results into fewer classes did improve the class accuracy, as expected. In fact, a binary munitions/not-munitions classification scheme yielded > 80% accuracy for most combinations of cameras and sites (Fig. 16). The remaining deficiencies, even with this reduced binary class scheme, were false positive matches for munitions in the reef and seagrass sites. The evidence for this was high producer's accuracy (all munitions were classified as munitions) but low user's accuracy (many pixels classified as munitions were actually something else).

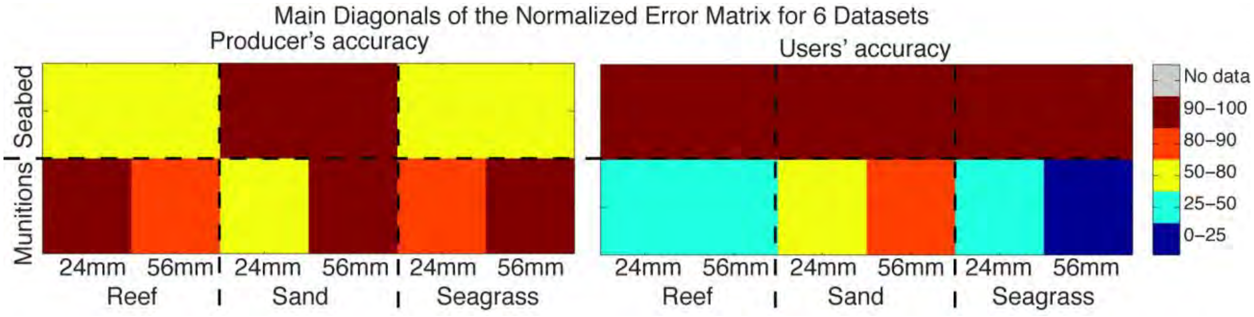


Figure 16. Main diagonals of consolidated versions of the normalized error matrix assessed with independent validation data for two cameras at all three of the Miami dataset sites: This figure is the same as Figure 15 except that the data have been collapsed from the full 36 classes to a binary scheme of munitions / background (Table 2). Note the high false positive rate in the reef and seagrass sites for identifying munitions characterized by high producer's and low user's accuracy.

In summary, the 2-D classification of the Miami dataset had excellent performance on its own training dataset, though less so on the validation dataset. Nevertheless, simple class schemes did well even on the validation dataset. The sand site, using the high-resolution (longer focal length) cameras, had > 84% accuracy with a binary munitions/not-munitions scheme, which was better than the same site classified using the lower resolution D7000 data (Fig. 16). The higher-resolution D7000 camera did not always produce better results, however (Figs 14 - 16). The weakest point of these results was an over-classification of munitions (too many false positives).

4.2.3 **Miami Dataset 2.5-D Results:** The same type of analysis used above, for the multiple camera, multiple site, Miami data also illuminated the differences between data processed using only 2-D information versus data processed using both the standard 2-D Girona features along with the extra 2.5-D information on height (Fig. 17). Like figures 14, 15, the columns of figure 17 correspond to main diagonals extracted from error matrices for different datasets, and the rows correspond to different classes defined in Table 2. Also like figures 14 and 15, separate panels of figure 17 contain information from the producer's and user's matrices. Unlike figures 14 and 15, however, Figure 17 has two versions each of the producer's user's images. One version corresponds to the results from the standard 2-D Girona algorithm whereas the other corresponds to the results from the modified 2.5-D algorithm. Note that independent evaluation data were used for all of these evaluations.

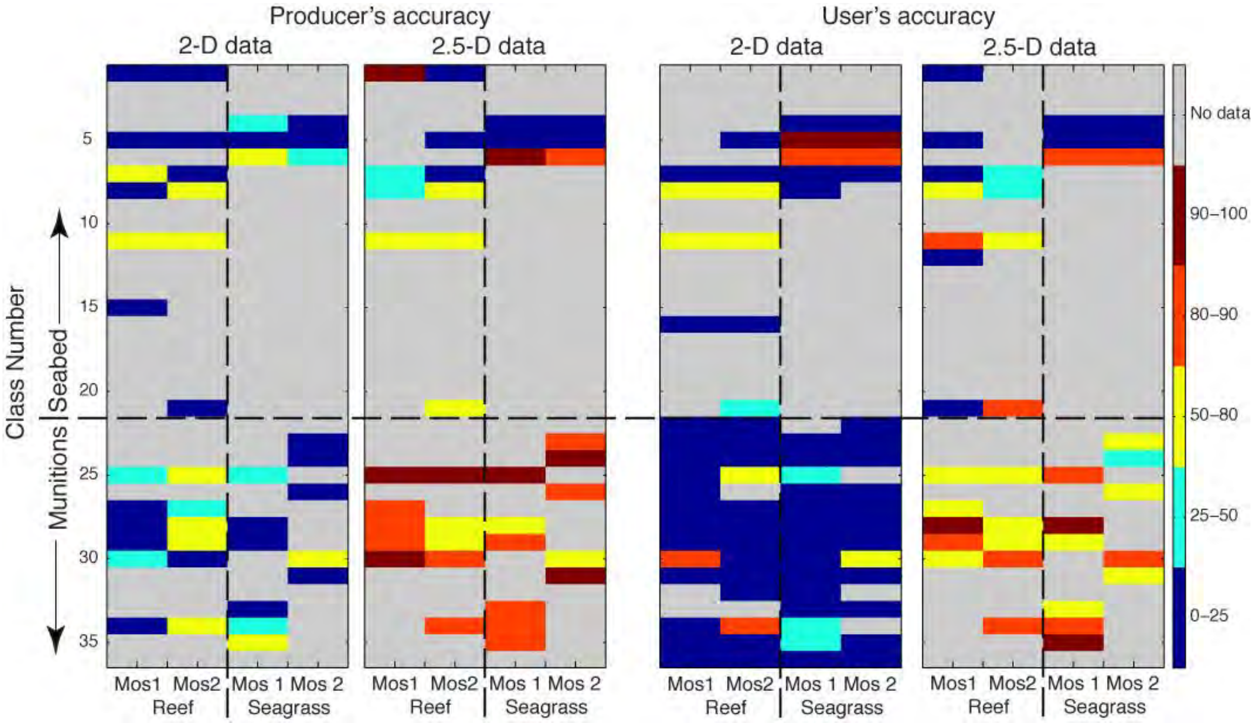


Figure 17. Main diagonals of the normalized error matrix assessed with independent validation data for four datasets processed with both 2-D and 2.5-D methods: Like Figures 14, 15 rows are defined by classes in Table 2. Columns correspond to one of 4 datasets that were processed with both 2-D and 2.5-D methods, 2 each in reef and seagrass habitats. The data processed with only 2-D algorithms had lower accuracy that the data processed with 2.5-D algorithms.

Adding the 2.5-D information did not change the results for the individual seabed classes very much, but it made a dramatic improvement in the accuracies of the munitions classes (Fig. 17). Most of the accuracies of the munitions classes using just 2-D were 0-80% range whereas these improved to 50-100% accuracy using the 2.5-D information (Fig. 17). Furthermore, adding the 2.5-D information improved the specific problem observed with the 2-D data, namely false positives for the munitions. This was evident in the error matrix by improved user's accuracies, relative to the strictly 2-D results. Note this improvement was observed even on a per-class basis, not just on the reduced binary munitions/not-munitions scheme.

Qualitatively, the improvement was evident on the imagery itself (Fig. 18). Results from the 2-D algorithm did classify the munitions correctly, but they overclassified the background as munitions (Fig. 18D). Resutls from the 2.5-D algorithm retained the high performance identifying even the types of individual munitions, but also greatly reduced the number of false positives (Fig. 18B, 18E).

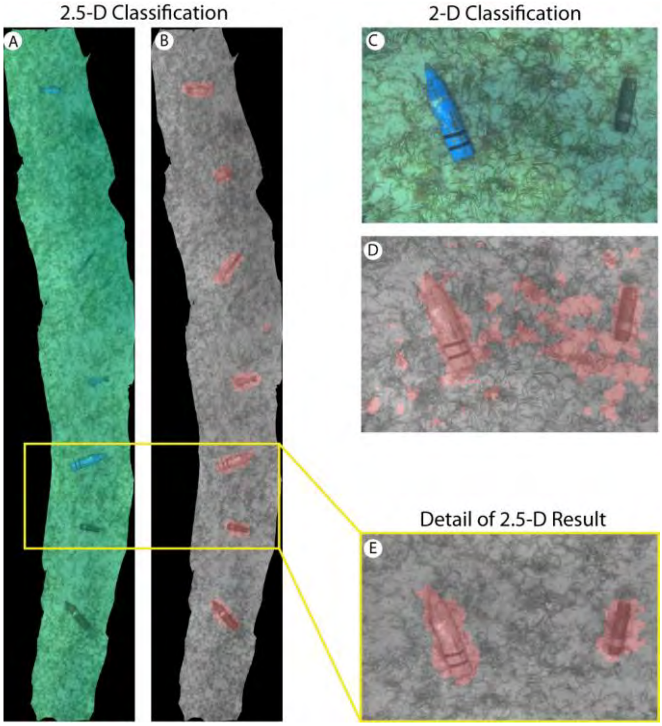


Figure 18. Qualitative comparison of images processed with both 2-D and 2.5-D methods: (A) Mosaic generated from ~20 individual, overlapping images, which was produced as part of the SfM process to derive height. (B) Greyscale version of (A) with areas classified as munitions highlighted in red. (C) One of the original images that was used to make (A). (D) A classified version of (C) produced using only 2-D methods. (E) Zoomed in portion of (B) for comparison with (D). Note that there are many fewer false positives in (E) than (D).

4.3 Dataset 3: HUMMA (Deep Water), Hawaii

Fewer, and different, classes were used for the HUMMA dataset (Table 1) than for the Miami dataset (Table 2). The results were analyzed in the same way, however. As for the Miami dataset, the Girona algorithm performed well on the HUMMA dataset when evaluated against the data used to train it, and less well when evaluated against independent validation data (Fig. 19). The accuracies for the HUMMA dataset with validation data, though low generally, did were not uniformly poor. The sand class, in particular, was very accurately classified even with the independent validation data (100% producer's, 83% user's accuracy). The confusion among the other classes (#2-7; Fig. 19) was mostly due to munitions being classified as "debris."

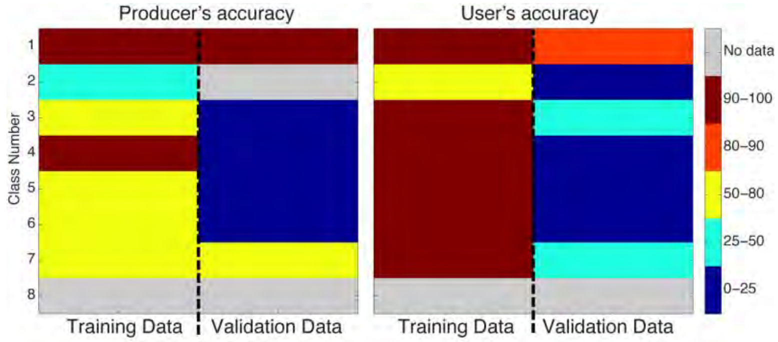


Figure 19. Main diagonals of the normalized error matrix assessed with the HUMMA dataset: Rows correspond to "all classes" defined in Table 1. The first column in both the left and right panel of this figure was produced from the training data used to define the HUMMA classification. The second column in both the left and right panel of this figure was produced from the independent validation data. Colorbar indicates accuracy as a percentage. Note the accuracy of all classes except sand (#1) declined substantially in the validation dataset relative to the training dataset.

Consolidating the HUMMA classes to a binary munitions/not-munitions scheme (Table 1) revealed a different pattern than was seen in the Miami data. The HUMMA data had high accuracy for the non-munitions classes but low producer's accuracy and high user's accuracy for the munitions classes (Fig. 20). The Miami dataset, in contrast, had high producer's and low user's accuracy for the munitions (Fig. 16). Thus, the HUMMA dataset underclassified munitions (too many false negatives) as opposed to the Miami dataset where munitions were overclassified.

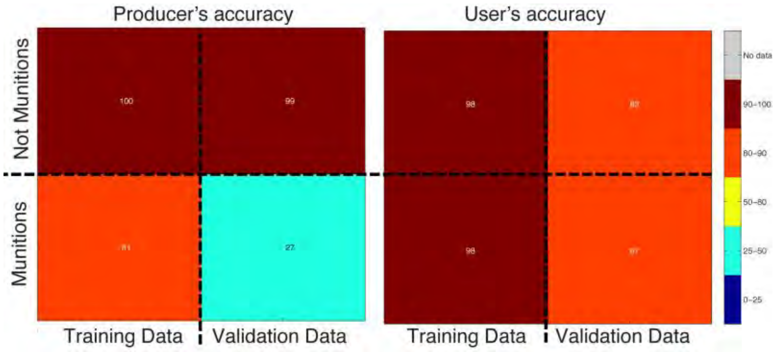


Figure 20. Main diagonals of the reduced normalized error matrix assessed with the HUMMA dataset: Rows correspond to "binary" classes defined in Table 1. The first column in both the left and right panel of this figure was produced from the training data used to define the HUMMA classification. The second column in both the left and right panel of this figure was produced from the independent validation data. Colorbar indicates accuracy as a percentage. Low producer's and high user's accuracy for munitions in the validation dataset indicates false negatives, or underclassification, of munitions.

Examining the validation images reinforced the interpretation derived from the error matrices. As expected from row 1 of Figure 19, the validation images confirmed that the background sand in the HUMMA images was accurately classified (Fig. 21). Looking just at the full-frame images (top row of Fig. 21) one can see that the areas classified as munitions or debris (shown in shades of red) were quite spatially homogenous, and not scattered across the backgound part of these images as had been observed in the Miami seagrass site (Fig. 18 were not classified as sand, which can be seen with the cropped versions of the images in the second row. The nature of the misclassification was also shown in the second row: many parts of the munitions were classified as “debris” of (Figs. 21 A, C, D). Note that actual debris was not usually classified as munitions, however (*e.g.* Fig. 21 B).

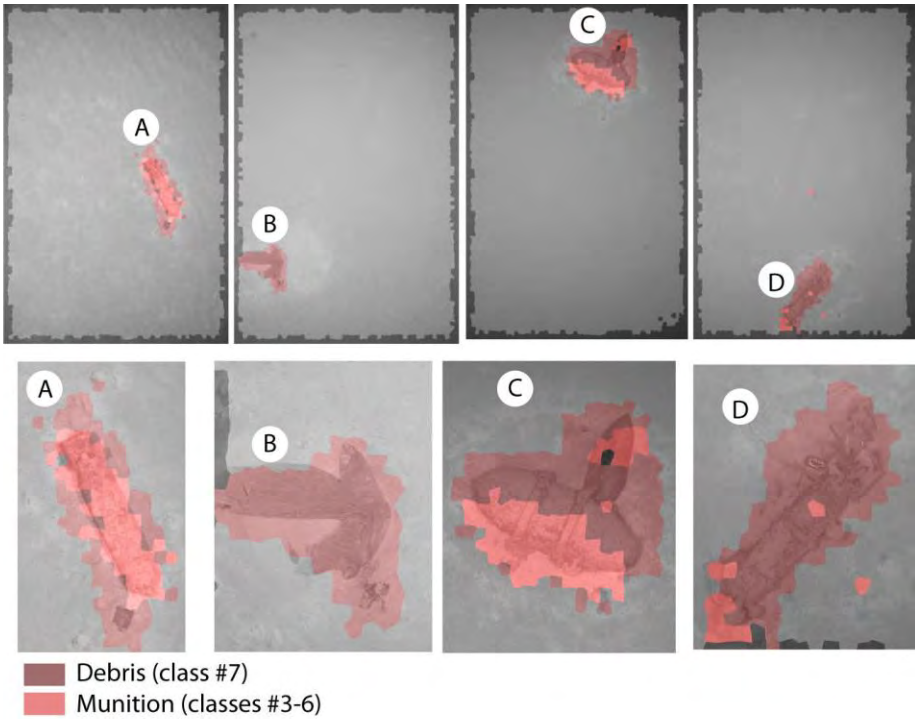


Figure 21. Example images from the HUMMA validation dataset: The top row shows four images from the independent validation dataset. The bottom row shows magnified portions of the corresponding images in the top row. Dark red pixels were classified as “debris” (Table 1, class #7). Light red pixels were classified as some type of munitions (Table 1, classes #3-6). Note that parts of munitions were often classified as debris, (A, C, D) but that debris was often classified correctly (B) and there were few to no false positives (high user’s accuracy for sand).

4.4 Image Classifier 2: WRT Features Applied to Miami Data

For each of the 63 WRT features, the distance between the mean (or median) value for the 311 background ROI samples and the mean (or median) value for the 63 target ROI samples was computed using four metrics: (1) Mann-Whitney and Wilcoxon Rank Sum, (2) Information Gain, (3) Mahalanobis and Fisher linear statistics, and (4) the Gini Index (Figure 22). The metrics represent simple univariate (or multivariate) feature selection filters that assess the discriminate power of each feature individually (and independently). The distance scores were correlated to a large degree with a few exceptions. These class distance metrics utilize different information about the feature set to derive scores and are not directly comparable. A more complete selection methodology should utilize rank aggregation or decision tree techniques in order to combine the metrics and create a master ranking. Nevertheless, after taking into account all distance measures, the eight features that had the largest average distances between the target and reference samples were the following (first 4 are marked on Fig. 22):

1. Sum of Covariances (Smoothness) [sumacov] feature #44
2. Sum of 1D Mean Values (Smoothness) [oneDmean] feature #51
3. Gray Level Homogeneity [GLHhomogeniety] feature #20
4. Cluster Prominence [clustprom] feature #11
5. Texture Feature Coding Method Code Entropy [TFCMSB2] feature #62
6. Texture Feature Coding Method Code Entropy [TFCMCV] feature #54
7. Number of Pixels Above 80% Max Threshold [NoPixelsGTthresj] feature #42
8. Along/Across Image Run Length [altrkLP] feature #35

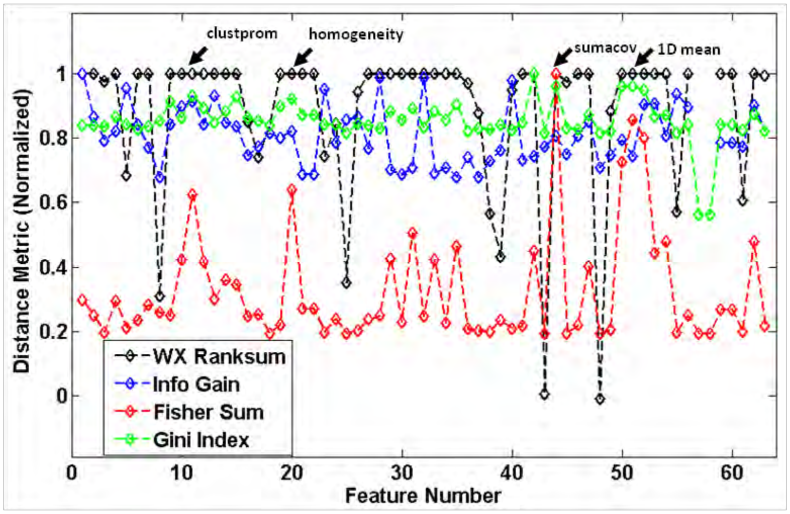


Figure 22. Normalized distance scores between target and background samples for each of the 63 WRT features relative to top performing feature (symmetry): The four distance metrics were Mann-Whitney and Wilcoxon Rank Sum (black), Information Gain (blue), Mahalanobis and Fisher linear statistics (red), and the Gini Index (green). Note that the Fisher sum distance metric seems to reveal several features that have larger average distances than the other features.

Inspecting the histograms of features that were shown to have large mean (or median) distances between the target points and background points showed that indeed the distributions of these two classes were quite different for widely separated features (Fig. 23). Most importantly, the features that had large distances between target points and background points accurately discriminated between the targets and background (Fig. 24). Overall accuracy discriminating targets from background using WRT feature #51, 1D Mean, and feature #44, sum of covariance, was 90%. Overall accuracy discriminating targets from background using WRT feature #42, number of pixels over the 80% threshold, and feature #22, roundness was also 90%.

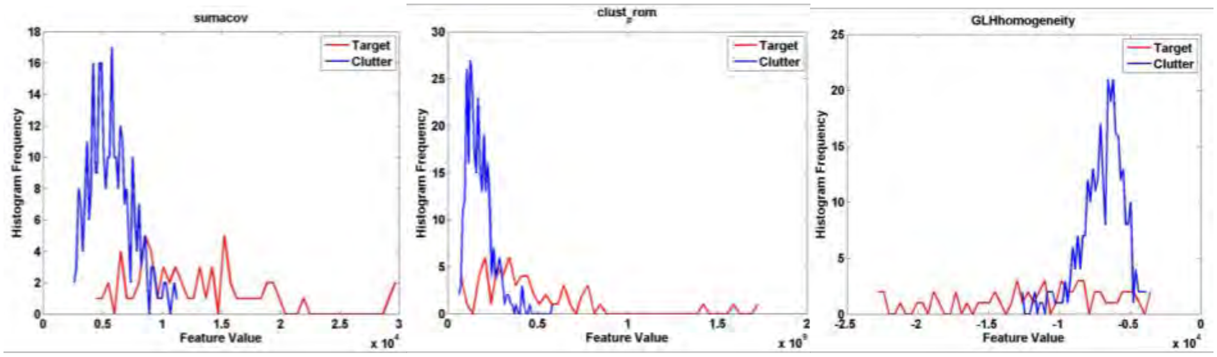


Figure 23. Histograms of single-feature values over 476 ROIs for three of the top performing features: Left: Sum of Covariances, feature #44 in Fig. 22. Center: Cluster Prominence, feature #11 in Fig. 22. Right: Gray Level Homogeneity, feature #20 in Fig. 22. Note that in each case, the distribution of target values for the feature (red line) was much different than the distribution of values for the background ROIs (blue line).

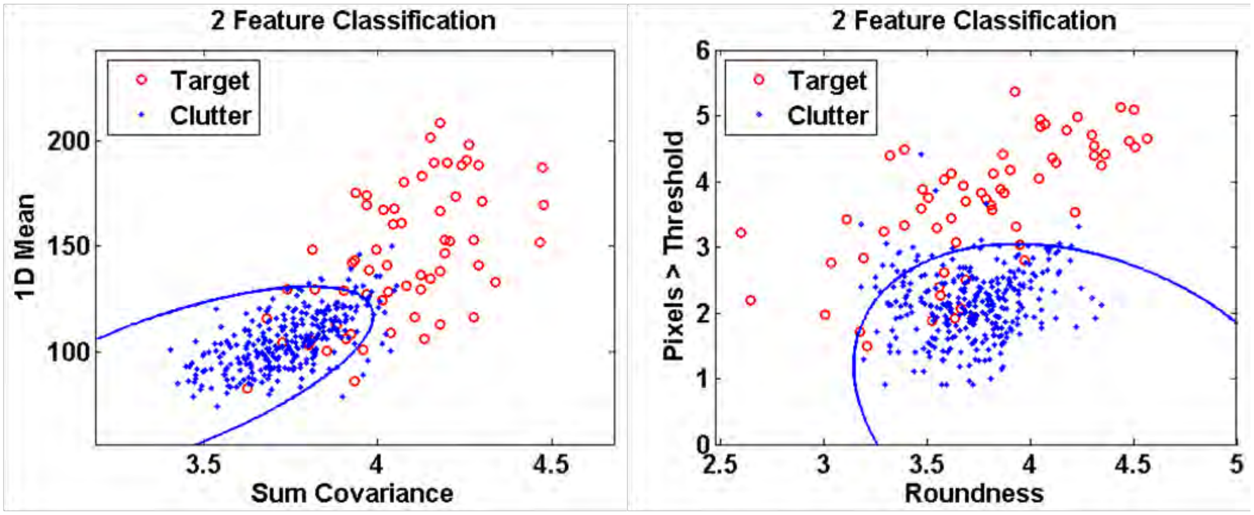


Figure 24. Quadratic discriminant functions applied to two pairs of WRT features: Left: WRT feature #51, 1D Mean, plotted against feature #44, sum of covariance. Right: WRT feature #42, number of pixels over the 80% threshold, plotted against feature #22, roundness. In both plots, red circles represent targets and blue dots represent background, or clutter, pixels. Quadratic discriminant functions have been plotted (blue lines), which minimize the probability of misclassifying a new feature vector. Note that the classes were generally well separated.

5 Conclusions and Implications for Future Research

The long-term goal of our research is to incorporate optics into a multi-modal approach to UWMM surveys. In order to accomplish this goal, automated analysis of large volumes of seabed imagery will be necessary to detect and ultimately classify munitions, discriminate munitions from clutter, and classify the surrounding seabed type. We identified a roadmap for developing such automated image analysis capability and for integrating imagery into surveys with other complementary instruments (Fig. 1). Before making the investment to follow that roadmap, however, it seemed prudent to evaluate the potential of our existing seabed classification algorithm in order to address questions such as: "is automated image analysis for munitions feasible at this time?" "is further algorithm development necessary?" "will fusion of multi-modal datasets complement the image-based approach?" Thus, the short-term goal of this project was to reduce the risk of investing in extensive algorithm improvement by demonstrating the performance of an existing seabed-image classification algorithm and the potential benefits of improving that algorithm, or one like it. Three objectives were accomplished to achieve this short-term goal.

Objective 1: Evaluate the potential for automated classification of underwater seabed images to improve both wide-area and detailed surveys for UWMM.

One result from this project was the indication that an existing seabed-image classification algorithm developed by our group does have promising potential to classify not just munitions vs. background clutter, but also to discriminate seabed types and even different types of munitions from one another. The image classifier by itself was shown to distinguish munitions from non-munitions (background) with generally high (> 80%) accuracy (Figs. 16, 20). This was accomplished at multiple sites in both shallow depths over seagrass, reef, and sand, and at depths greater than 500 m in sand.

The image classifier by itself did not always achieve > 80% accuracy, however, even with a binary munitions/non-munitions scheme. In the Miami dataset, false positives were observed over reef and seagrass (Fig. 16), whereas in the HUMMA dataset, false negatives were observed due to confusion of munitions with other anthropogenic clutter (Fig. 20). These limitations indicated that there is indeed a need for improvements to the algorithm.

Objective 2: Evaluate the potential for automated classification of underwater seabed images to improve characterization of the environment surrounding UWMM by discriminating bottom types in general (rock, sediment, seagrass etc..) as well as identifying specific coral species.

Discrimination of general seabed types was high for the major seabed types. For example, sand and mixed sand-seagrass were classified with 80-100% accuracy in both the Miami and HUMMA datasets, even when evaluated using independent validation data (Figs. 14**,** 15**,** 19). Coral, even at the species level, was also classified with > 80% accuracy in the training dataset. Due to the low coral cover at the test sites (~3%). there were not enough samples of corals in the randomly selected validation dataset to evaluate these classes with independent validation data. Performance of the classifier on the training data was similar for coral species and munitions types (Fig. 14), and the increased accuracy for the coral class when using 2.5-D data vs. 2-D data only was also similar to that observed for the munitions classes (Fig. 17).

As was the case for the munitions classes, the Girona algorithm by itself did not always achieve > 80% accuracy for all seabed classes. Some of the errors in seabed classification were the converse of those observed for munitions. Namely, the munitions were overclassified (false positives), so by definition, the seabed classes were underclassified (false negatives). On the other hand, some of the errors among the seabed classes were due to the "fuzzy" nature of the boundaries between some natural classes. "Seagrass", "sand and seagrass", and "sand", for example are three class labels that divide what is really a continuous gradient from 100% seagrass to 100% bare sediment. Likewise, the difference between turf algae, crustose-coralline algae, bare substrate, and sand in the reef site are also gradations along a continuum. Distinguishing among classes with fuzzy boundaries like these is a challenge even for human analysts. Beijbom et al. (in review), for example, found that inter- and intra-observer variability was much higher for turf and crustose-coralline algae than for coral genera.

Objective 3: Test the assumption that further improvements to the current generation of seabed classification algorithms will yield further improvements in the ability to discriminate munitions and aspects of the environment, such as coral species identification.

In order to test whether classification accuracy could be improved with further algorithm development, we tested one of many ideas we have for extending the existing algorithm. The extension that was developed used stereo reconstruction derived from overlapping views of the same area of seabed to generate additional features based on local relief. The results showed that incorporating such so-called "2.5-D" data greatly improved the classification results (Figs. 17, 18). Using the 2.5-D information reduced the number of false positives for munitions in the Miami dataset, increasing accuracy from the 0-80% range to the 50-100% range (Fig. 17). Furthermore, these improved accuracies were observed not only on the basic, binary munitions / non-munitions classes; adding 2.5-D information improved the capability to discriminate different types of munitions from one another. Improvements gained by adding one extra source of information (2.5-D data) suggest that adding other additional data, such as object morphology derived from the images and other data derived from complementary sensors, would be a logical next step for further accuracy improvement.

Another result of this project was to eliminate some potential future development paths. It was shown that the Girona algorithm and the WRT features, a different set of 2-D image features, achieved similar overall accuracies. Thus, we conclude that adding more statistical or texture-based 2-D features to the algorithm, while always possible, probably would not result in large gains in accuracy. This result was useful because it is just as valuable to eliminate potentially unproductive development paths as it is to conceive of new ones.

Next steps: This SEED project demonstrated that automated classification of underwater seabed images has the potential to improve both wide-area and detailed surveys for UWMM by identifying munitions themselves as well as characterizing the environments surrounding UWMM. Automated classification at high levels (*e.g.* munitions types or coral species) was possible, though much more robust when the 2.5-D algorithm was used than when the strictly 2-D algorithm was used. Therefore, we expect that further improvements to the current generation of seabed classification algorithms will yield improvements in the ability to discriminate munitions and aspects of the environment, and suggest that this should be the next step towards incorporating underwater imagery into a multi-modal approach to UWMM surveys

Based on the WRT features and 2.5-D experiments, the most promising way forward will add fundamentally different types of data to the basic color and 2-D texture features. Adding 2.5-D or shape features or data from other complementary sensors are different than just adding new 2-D texture features because the former is really fusing new information into the problem as opposed to the latter, which consists of just processing the same existing data in different ways. Two new objectives that are logical next steps given the results of this SEED proposal are:

Next Objective 1: Improve the automated detection and discrimination of underwater military munitions (UWMM) by fusing optical features derived from underwater images (in particular shape, 2.5-D relief, and image texture) with magnetometer and acoustic datasets.

Next Objective 2: Improve the automated detection and discrimination of benthic organisms, specifically corals, by incorporating shape and multi-scale features into optical seabed classification algorithms.

As originally outlined in the gray box of Figure 1, we continue to feel that the greatest practical use of seabed imagery for UWMM mapping and remediation is automated processing of image data that is synchronized/co-located with sonar or magnetics data. Fusing optical imagery with these other data types would improve the performance of classifications based on the suite of data relative to using any one data source alone. An efficient next step would addresses objectives 1 and 2 in parallel, thereby allowing for progress toward objective 2 over the duration of the project to "feed back" by also improving objective 1 (Fig. 1). The common starting point for both objectives will be the existing Girona algorithm with the 2.5-D extensions. The approach for objective 1 will start with the existing optical algorithm as-is, and combine it with classification methods for complementary acoustic and magnetic datasets. The approach for objective 2 will be to improve on the existing optical algorithm by incorporating new features based on shape and scale. By building on the base of existing underwater characterization technologies, we envision that the proposed multi-modal approach will yield similar benefits to those realized under previous terrestrial wide area assessment projects.

Literature Cited

Atallah, L. and P. J. P. Smith (2004). Automatic seabed classification by the analysis of sidescan sonar and bathymetric imagery. *Radar, Sonar and Navigation, IEE Proceedings -* **151**(5): 327-336.

Beijbom, O., P. J. Edmunds, D. I. Kline, B. G. Mitchell and D. Kriegman (2012). Automated annotation of coral reef survey images. *Proceedings of the 2012 IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, 16-21 June 2012: pp. 1170-1177.

Beijbom, O., P. J. Edmunds, C. M. Roelfsema, J. Smith, D. I. Kline, B. Neal, M. J. Dunlap, V. Moriarty, T.-Y. Fan, C.-J. Tan, S. Chan, T. Treibitz, A. Gamst, B. G. Mitchell and D. Kriegman (in review). Towards automated annotation of benthic survey images: variability of human experts and operational modes of automation. *PLoS ONE*.

Congalton, R. G. and K. Green (1999). Assessing the Accuracy of Remotely Sensed Data: Principles and Practices, Boca Raton, Lewis Publishers, 137 pp.

Dell'Acqua, F. and P. Gamba (2003). Texture-based characterization of urban environments on satellite SAR images. *IEEE Transactions on Geoscience and Remote Sensing* **41**(1): 153-159.

Edwards, M. H., R. Wilkens, C. Kelley, E. DeCarlo, K. MacDonald, S. Shjegstad, M. V. Woerkom, Z. Payne, V. Dupra, M. Rosete, M. Akiba, S. Fineran, W. Zheng, J. C. King and G. Carton (2012). Methodologies for Surveying and Assessing Deep-Water Munitions Disposal Sites. *Marine Technology Society Journal* **46**(1): 51-62.

Finn, J. T. (1993). Use of the Average Mutual Information Index in Evaluating Classification Error and Consistency. *International Journal of Geographical Information Systems* **7**(4): 349-366.

Foley, J. and J. Hodgson (2010). Wide area assessment - Development and case study. *Proceedings of the 2010 IEEE International Geoscience and Remote Sensing Symposium (IGARSS)*, 25-30 July 2010: pp. 3370-3373.

Fornari, D. J. and The WHOI Towcam Group (2003). A New Deep-sea Towed Digital Camera and Multi-rock Coring System. *Eos* **84**(8): 69-76.

Foster, G., A. Gleason, B. Costa, T. Battista and C. Taylor (2013). Acoustic Applications. in Coral Reef Remote Sensing. J. A. Goodman, S. J. Purkis and S. R. Phinn, Eds., Springer Netherlands**:** pp. 221-251.

Friedman, A., O. Pizarro and S. Williams (2010). Rugosity, slope and aspect from bathymetric stereo image reconstructions. *Proceedings of the IEEE Oceans Conference, OCEANS 2010*, Sydney, Australia: pp. 1-9.

Friedman, A., O. Pizarro, S. Williams and M. Johnson-Roberson (2012). Multi-Scale Measures of Rugosity, Slope and Aspect from Benthic Stereo Image Reconstructions. *PLoS ONE* **7**(12).

Haralick, R. M., Shanmuga.K and I. Dinstein (1973). Textural Features for Image Classification. *IEEE Transactions on Systems, Man, and Cybernetics* **3**(6): 610-621.

Hetzel, G., B. Leibe, P. Levi and B. Schiele (2001). 3D object recognition from range images using local feature histograms. *Proceedings of the IEEE Computer Society Conference on Computer Vision and Pattern Recognition (CVPR)*: pp. 394-399.

Ho, K., L. Carin, P. Gader and J. Wilson (2008). An investigation of using the spectral characteristics from ground penetrating radar for landmine/clutter discrimination. *IEEE Transactions on Geoscience and Remote Sensing* **46**: 1177-1191.

Horng, M.-H., Y.-N. Sun and X.-Z. Lin (2002). Texture feature coding method for classification of liver sonography. *Computerized Medical Imaging and Graphics* **26**(1): 33-42.

Kelley, C., G. Carton, M. Tomlinson and A. Gleason (in press). Analysis of towed camera images to determine the effects of disposed mustard-filled bombs on the deep water benthic community off south Oahu. *Deep Sea Research Part II: Topical Studies in Oceanography*.

Keranen, J. and J. O'Neil-Dunne (2010). Combining Synthetic Aperture Radar (SAR) And Light Detection And Ranging (LiDAR) For Advanced Improvised Explosive Devise (IED) Detection And Intelligence, Surveillance, And Reconnaissance (ISR) Applications. SBIR Phase 1 Report AFRL/RYAR: AFRL-RY-WP-TR-2010-1299, 77 pp.

Keren, D. and C. Gotsman (1999). Fitting curves and surfaces with constrained implicit polynomials. *IEEE Transactions on Pattern Analysis and Machine Intelligence* **21**(1): 31-41.

Kohler, K. E. and S. M. Gill (2006). Coral Point Count with Excel extensions (CPCe): A Visual Basic program for the determination of coral and substrate coverage using random point count methodology. *Computers and Geosciences* **32**(9): 1259-1269.

Kovesi, P. (1997). Symmetry and asymmetry from local phase. *Proceedings of the Tenth Australian Joint Convergence on Artificial Intelligence*: pp. 2-4.

Kreithen, D. E., S. D. Halversen and G. J. Owirka (1993). Discriminating targets from clutter. *The Lincoln Laboratory Journal* **6**: 25-52.

Li, Q., J. Ye and C. Kambhamettu (2003). Spatial interest pixels: Useful low-level features of visual media data. *Proceedings of the 3rd IEEE International Conference on Data Mining (ICDM 2003)*, Melbourne, Florida, 19-22 December 2003, IEEE Computer Society: pp. 163-170.

Lirman, D., N. R. Gracias, B. E. Gintert, A. C. R. Gleason, R. P. Reid, S. Negahdaripour and P. Kramer (2007). Development and application of a video-mosaic survey technology to document the status of coral reef communities. *Environmental Monitoring and Assessment* **125**: 59-73.

Marcos, M. S. A., L. David, E. Penaflor, V. Ticzon and M. Soriano (2008). Automated benthic counting of living and non-living components in Ngedarrak Reef, Palau via subsurface underwater video. *Environmental Monitoring and Assessment* **145**(1-3): 177-184.

Nakatsu, J. S. K., H.-S. Youn and M. Iskander (2012). Feasibility study for non-metallic IED detection using forward-looking ground penetrating radar integrated with target feature classification. *IEEE Antennas and Propagation Society International Symposium (APSURSI)*, Chicago, IL, 8-14 July 2012: pp. 1-2.

Nelson, H., K. Kaye and A. Andrews (2008). ESTCP Pilot Project Wide Area Assessment for Munitions Response, Environmental Security Technologies Certification Program, 340 pp.

Pican, N., E. Trucco, M. Ross, D. M. Lane, Y. Petillot and I. T. Ruiz (1998). Texture analysis for seabed classification: co-occurrence matrices vs. self-organizing maps. *Proceedings of the IEEE Oceans Conference (OCEANS '98)*, 28 Sep-1 Oct 1998: pp. 424-428.

Pizarro, O., P. Rigby, M. Johnson-Roberson, S. B. Williams and J. Colquhoun (2008). Towards image-based marine habitat classification. *Proceedings of the IEEE Oceans Conference (OCEANS 2008)*, 15-18 Sept. 2008: pp. 1-7.

Schechner, Y. Y. and N. Karpel (2005). Recovery of underwater visibility and structure by polarization analysis. *IEEE Journal of Oceanic Engineering* **30**(3): 570-587.

Schmid, C., R. Mohr and C. Bauckhage (2000). Evaluation of interest point detectors. *International Journal of Computer Vision* **37**: 151-172.

Schultz, G. M., J. Foley and S. Billings (2009). Finding Underwater Munitions: Technologies and Applications. *Sea Technology* **53**(3): 19-24.

Schwartz, A. and E. Brandenburg (2009). An Overview of Underwater Technologies for Operations Involving Underwater Munitions. *Marine Technology Society Journal* **43**(4): 62-75.

SERDP and ESTCP (2007). Final Report: SERDP and ESTCP Workshop on Technology Needs for the Characterization, Management, and Remediation of Military Munitions in Underwater Environments, San Diego, CA July 31-Aug 1, 2007, 45 pp.

Shanmugan, K. S., V. Narayanan, V. S. Frost, J. Abbot and J. Holtzman (1981). Texture features for radar image analysis. *IEEE Transactions on Geoscience and Remote Sensing* **19**: 153-156.

Shihavuddin, A. S. M., N. Gracias, R. Garcia, R. Campos, A. C. R. Gleason and B. Gintert (2014). Automated Detection of Underwater Military Munitions Using Fusion of 2-D and 2.5-D Features from Optical Imagery. *Marine Technology Society Journal* **48**(4): 61-71.

Shihavuddin, A. S. M., N. Gracias, R. Garcia, A. Gleason and B. Gintert (2013). Image-Based Coral Reef Classification and Thematic Mapping. *Remote Sensing* **5**(4): 1809-1841.

Stokes, M. D. and G. B. Deane (2009). Automated processing of coral reef benthic images. *Limnology and Oceanography-Methods* **7**: 157-168.

Turcotte, D. L. (1997). Fractals and Chaos in Geology and Geophysics. 2nd Edition, Cambridge University Press

Ulaby, F. T., B. F. B. Kouyate and L. Williams (1986). Textural information in SAR images. *IEEE Transactions on Geoscience and Remote Sensing* **24**: 235-245.

Wang, Y., B. S. Peterson and L. H. Staib (2000). Shape-based 3-D surface correspondence using geodesics and local geometry. *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*: pp. 644-651.

Appendices:

Enclosed as an appendix is a copy of Shihavuddin, et al. (2014), which was written with support from this project.